Fast campus

2017 Machine Learning with R

Multiple Linear Regression

강필성

고려대학교 산업경영공학부

pilsung_kang@korea.ac.kr

목차

I 다중선형회귀분석

II 회귀분석 성능 평가

III 변수 선택

IV 다중선형회귀분석 예시

V R 실습

회귀분석

❖ 도요타 코롤라 자동차의 중고차 가격 예측

종속 변수
(target)

설명 변수
(attributes, features)

Variable	Description
Price	Offer Price in EUROs
Age_08_04	Age in months as in August 2004
KM	Accumulated Kilometers on odometer
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)
HP	Horse Power
Met_Color	Metallic Color? (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters
Doors	Number of doors
Quarterly_Tax	Quarterly road tax in EUROs
Weight	Weight in Kilograms

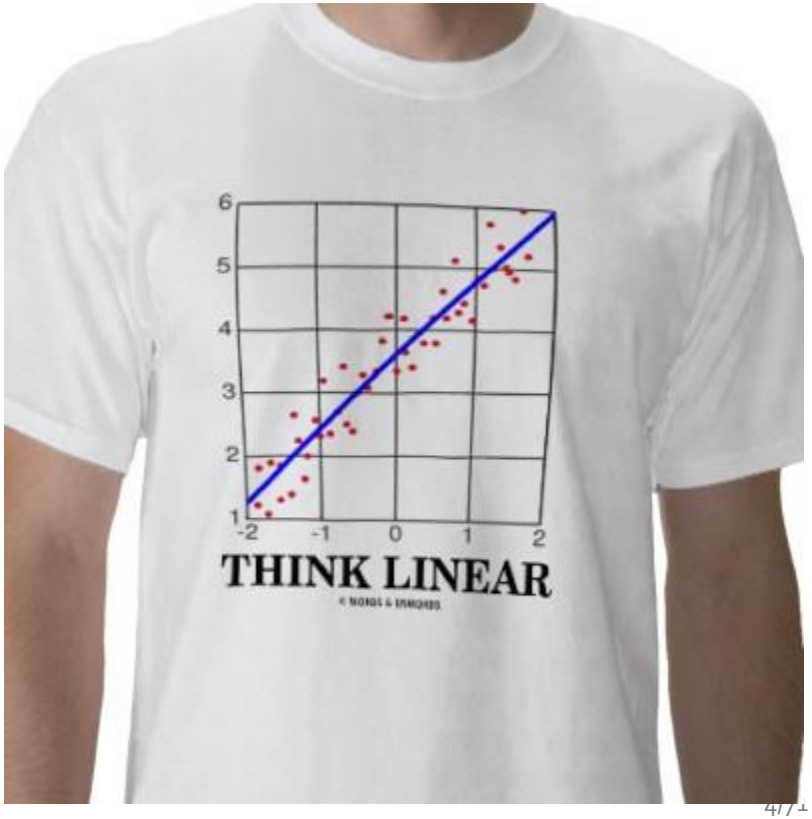
다중회귀분석: 목적

❖ 목적

- 종속변수 Y와 설명변수 집합 X_1, X_2, …, X_p사이의 관계를 선형으로 가정하고 이를 가장 잘 설명할 수 있는 회귀 계수(regression coefficients)를 추정

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

coefficients　　unexplained



다중회귀분석: 탐색적 vs. 예측적

❖ 탐색적(Explanatory) 회귀분석 vs. 예측적(Predictive) 회귀분석

Explanatory Regression

- Explain relationship between predictors (explanatory variables) and target.
- Familiar use of regression in data analysis.
- Model Goal: Fit the data well and understand the contribution of explanatory variables to the model.
- "goodness-of-fit": R^2, residual analysis, p-values.

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

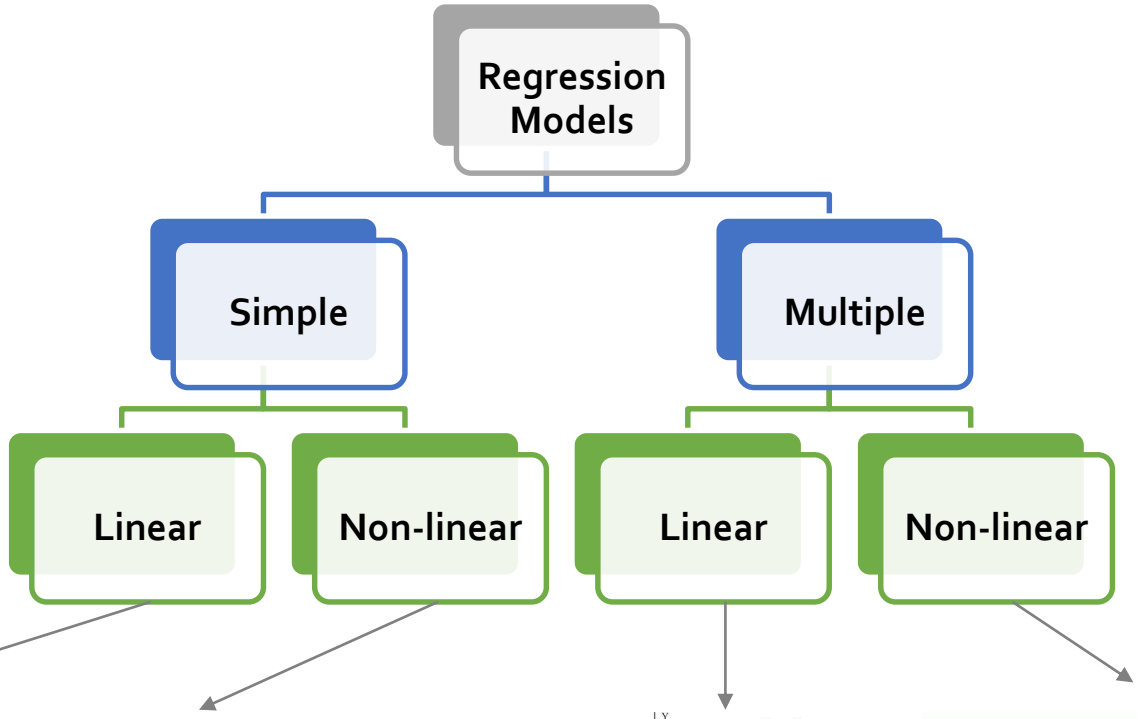
Predictive Regression

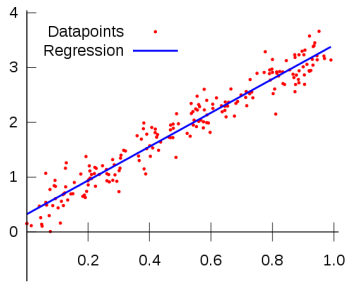
- Predict target values in other data where we have predictor values, but not target values.
- Classic data mining context
- Model Goal: Optimize predictive accuracy
- Train model on training data
- Assess performance on validation (hold-out) data
- Explaining role of predictors is not primary purpose (but useful)

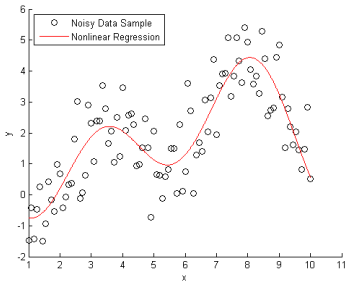
$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

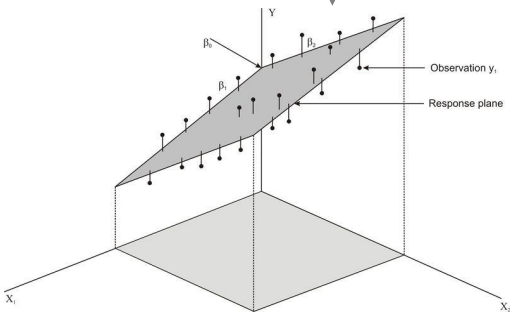
회귀분석의 형태

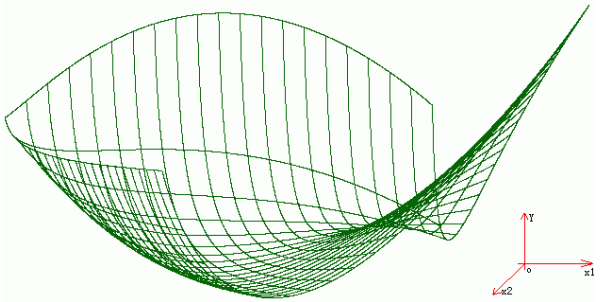
❖ 변수의 수와 추정되는 함수의 형태에 따라 아래와 같이 구분 가능









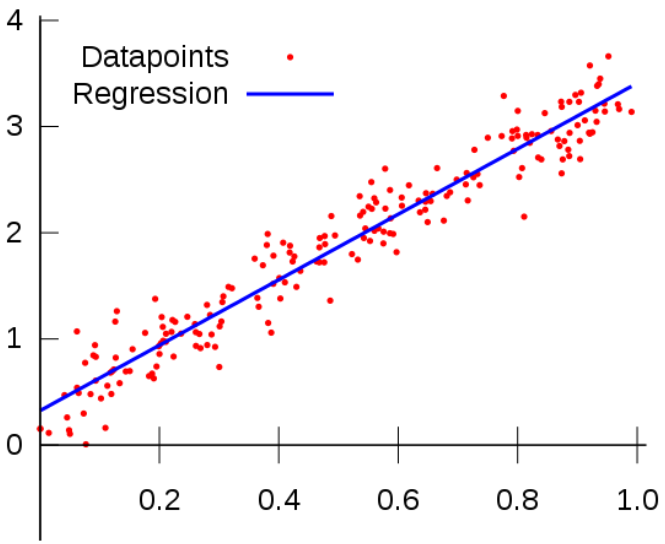


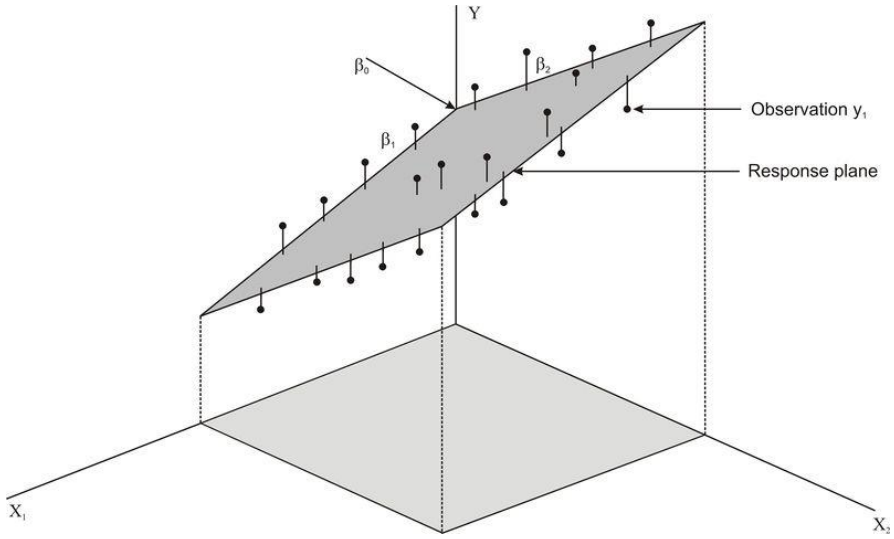
선형 회귀분석

❖ 선형 회귀 분석: Linear Regression

- 반응변수와 설명변수 사이의 관계를 선형으로 표현

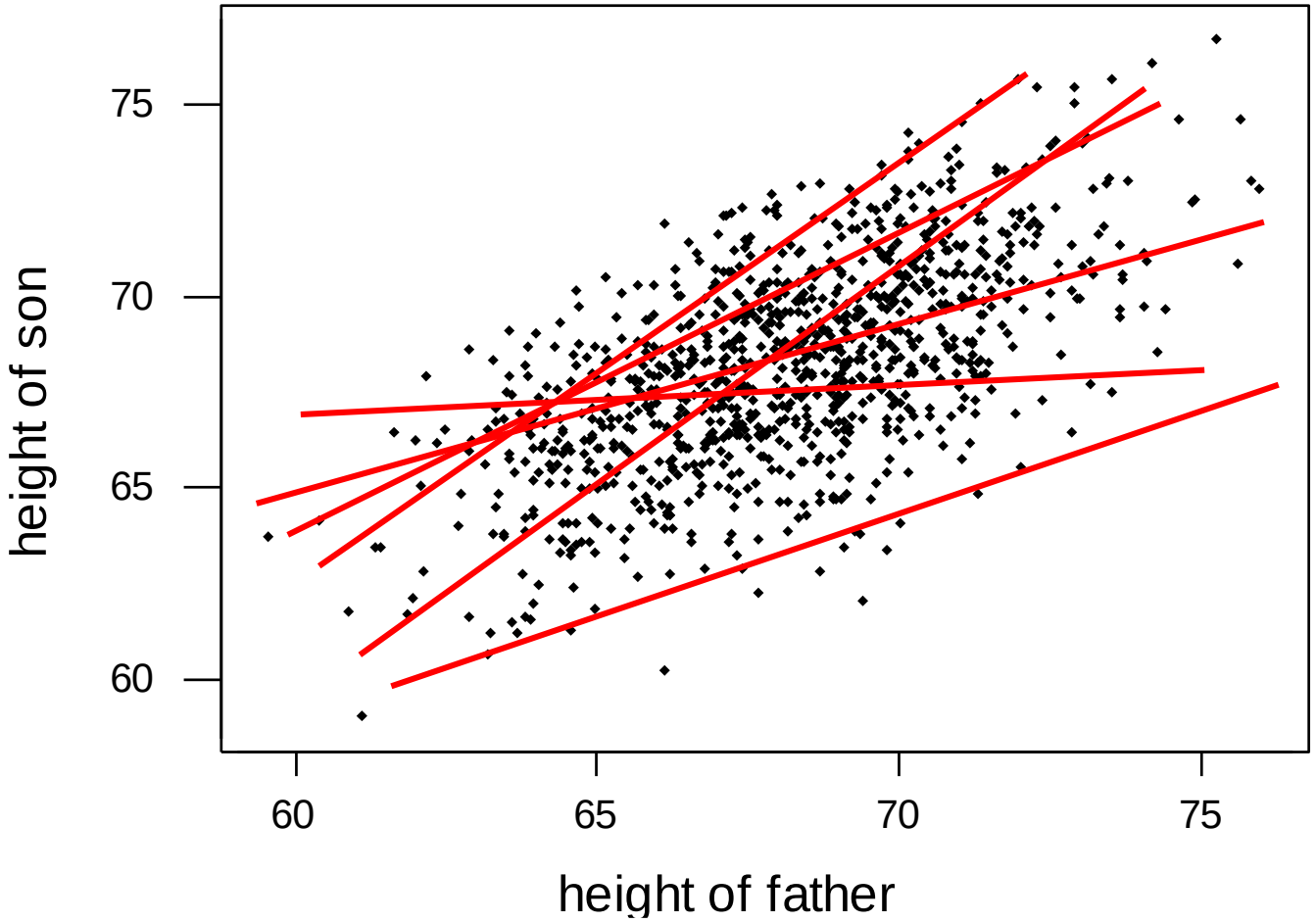
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$





선형 회귀분석

❖ 어떤 직선이 설명변수와 종속변수를 가장 잘 표현하는가?



다중회귀분석: 회귀 계수의 추정

❖ 회귀 계수의 추정

- 최소자승법: Ordinary least square (OLS)
  - ✓ Actual target: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$
  - ✓ Predicted target: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$
  - ✓ 목적: 실제 종속변수 값과 예측된 종속변수 값 사이의 오차 제곱항을 최소화

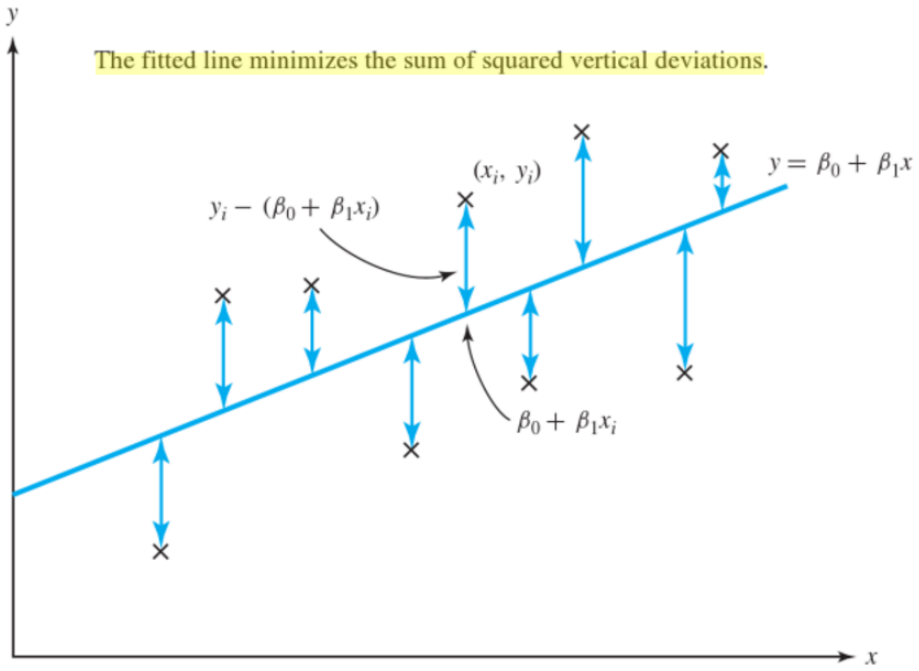
$$\min \quad \frac{1}{2}\sum_{i=1}^{N} \varepsilon_i^2 = \frac{1}{2}(Y_i - \hat{Y}_i)^2$$

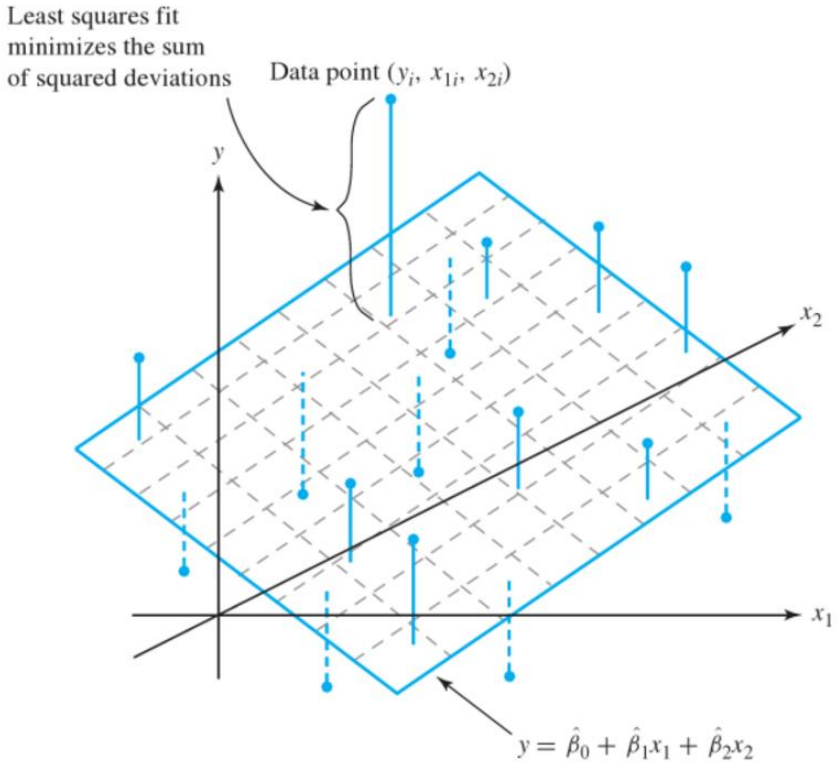
$$= \frac{1}{2}\left(Y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \ldots - \hat{\beta}_p x_p\right)^2$$

다중회귀분석: 회귀 계수의 추정

❖ 회귀 계수의 추정

- 최소자승법: Ordinary least square (OLS)





다중회귀분석: 회귀 계수의 추정

❖ 최소 자승법: 행렬을 이용한 해 구하기

- **X**: n by p matrix, **y**: n by 1 vector, **β**: p by 1 vector.

$$\min \quad E(\mathbf{X}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\Rightarrow \quad \frac{\partial E(\mathbf{X})}{\partial \boldsymbol{\beta}} = -\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

$$\Rightarrow \quad -\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = 0$$

$$\Rightarrow \quad \boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

다중회귀분석: 회귀 계수의 추정

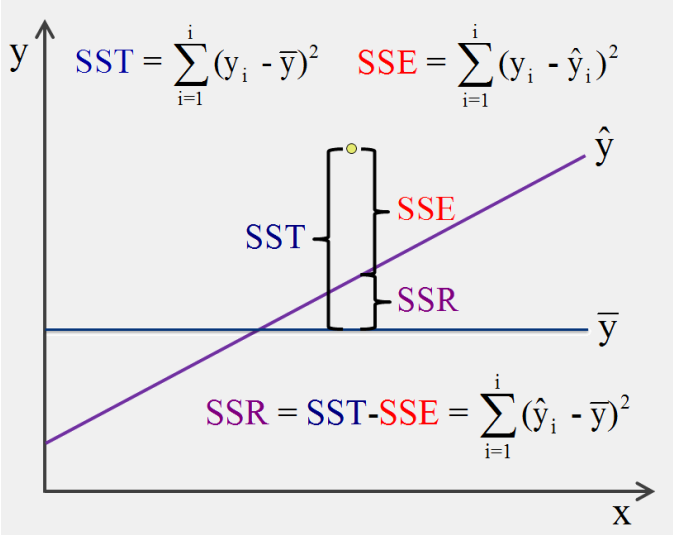
❖ 최소자승법

- 아래 조건을 만족할 경우 최소자승법으로 구한 회귀계수 β는 최적해임
  - 오차항 ε 이 정규분포를 따름
  - 설명변수와 종속변수 사이에 선형관계가 성립함
  - 각 관측치들은 서로 독립
  - 종속변수 Y에 대한 오차항(residual)은 설명변수 값의 범위에 관계없이 일정함(homoskedasticity)

다중회귀분석: 회귀 계수의 추정

❖ 회귀모형의 적합도

- Sum-of-Squares Decomposition



$$\underbrace{\sum_{j=1}^{n}(y_j-\bar{y})^2}_{\substack{\text{total sum of squares}\\ \text{about mean}}} = \underbrace{\sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2}_{\substack{\text{regression}\\ \text{sum of squares}}} + \underbrace{\sum_{j=1}^{n}\hat{\varepsilon}_j^2}_{\substack{\text{residual (error)}\\ \text{sum of squares}}}.$$

SST **SSR** **SSE**

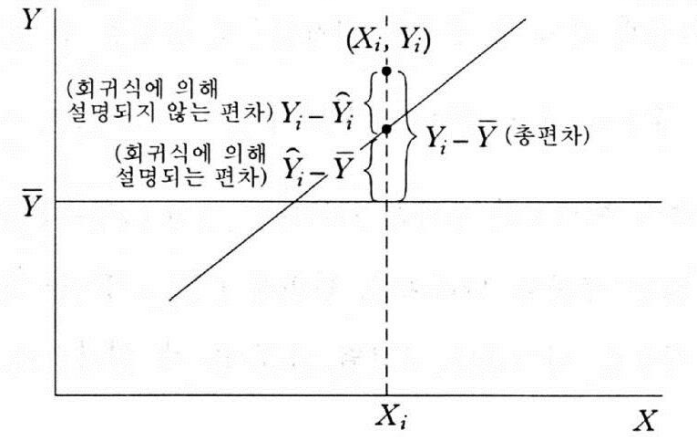
다중회귀분석: 회귀 계수의 추정

❖ 회귀모형의 적합도

- 결정계수(R^2):

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

$$0 \leq R^2 \leq 1$$



✓반응변수 (Y)의 전체 변동 중 예측변수(X)가 차지하는 변동의 비율

✓R^2는 0과 1사이에 존재

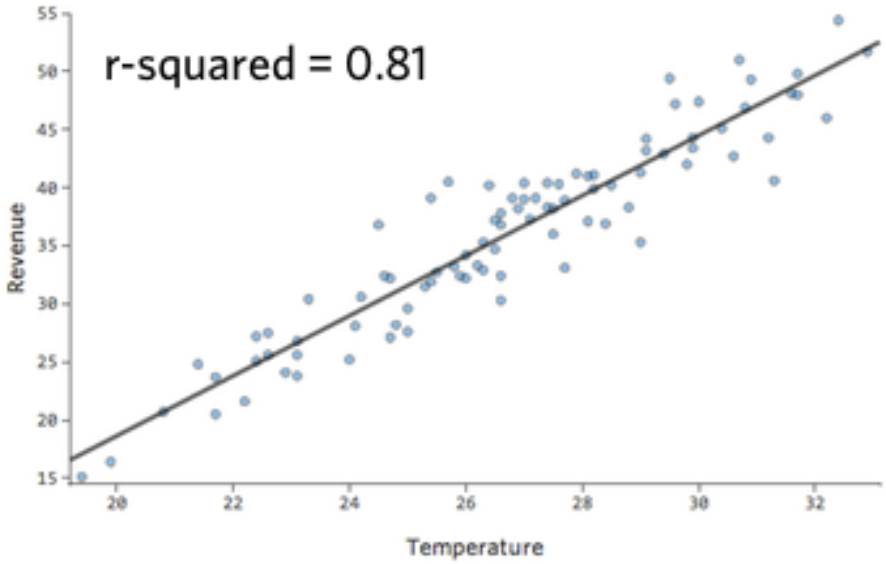
✓R^2=1: 회귀직선으로 Y의 총변동이 완전히 설명됨 (모든 측정값들이 회귀직선 위에 있는 경우)

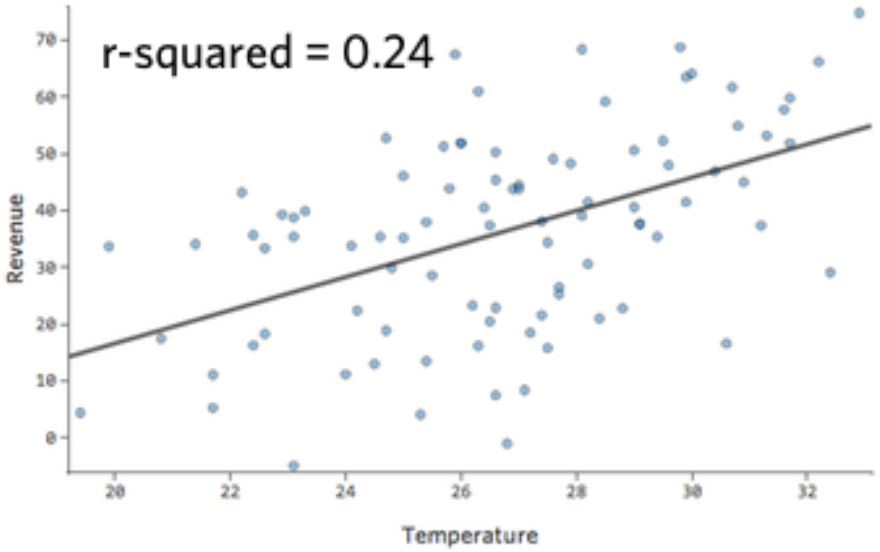
✓R^2=0: 추정된 회귀직선은 X와 Y의 관계를 전혀 설명하지 못함

다중회귀분석: 회귀 계수의 추정

❖ 회귀모형의 적합도

- 결정계수(R^2):





다중회귀분석: 회귀 계수의 추정

❖ 회귀모형의 적합도

- 수정 결정계수(Adjusted R^2):

$$R^2_{adj} = 1 - \left[\frac{n-1}{n-(p+1)}\right]\frac{SSE}{SST} \leq 1 - \frac{SSE}{SST} = R^2$$

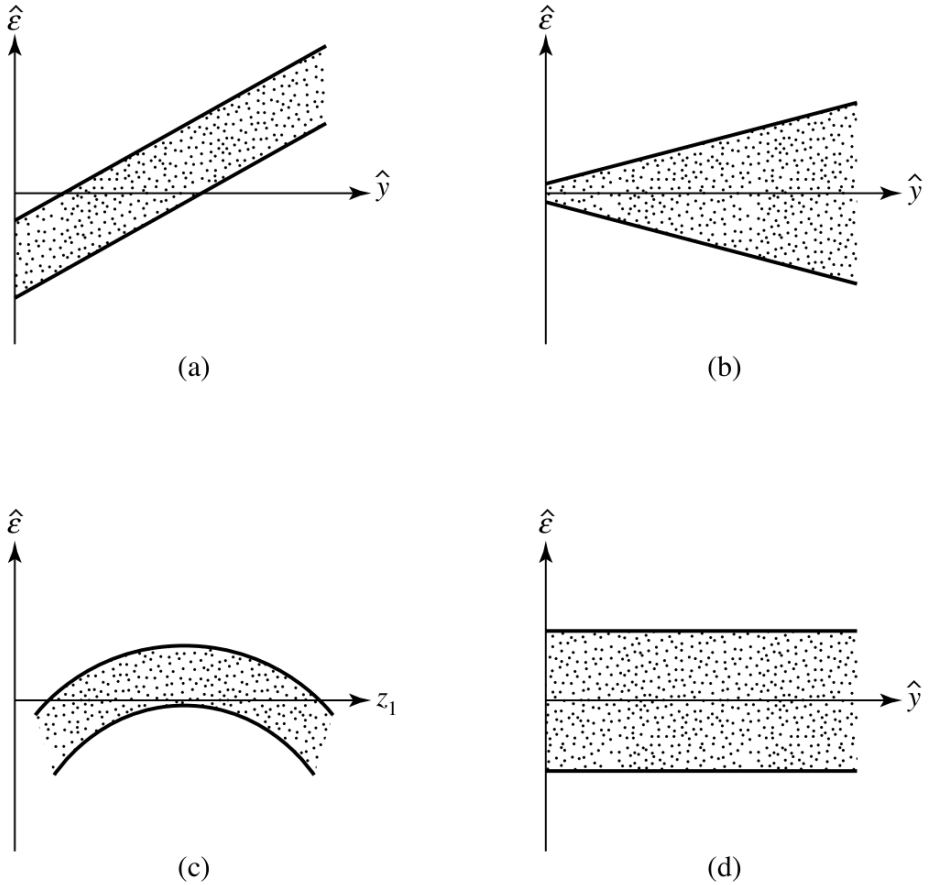
  - ✓ R^2는 유의하지 않은 변수가 추가되어도 항상 증가
  - ✓ 수정 R^2는 이러한 단점을 앞에 계수를 곱해줌으로써 보정
  - ✓ 유의하지 않은 변수가 추가될 경우 수정 결정계수는 증가하지 않음

❖ 모형의 검토

- 추정된 모형이 다음 가정을 만족하는지 확인
  - ✓ 예측변수와 반응변수 간 관계가 선형
  - ✓ 오차항들이 서로 독립
  - ✓ 오차항은 평균이 0이며 분산이 일정한 정규분포를 따름

다중회귀분석: 모형의 적합도 평가

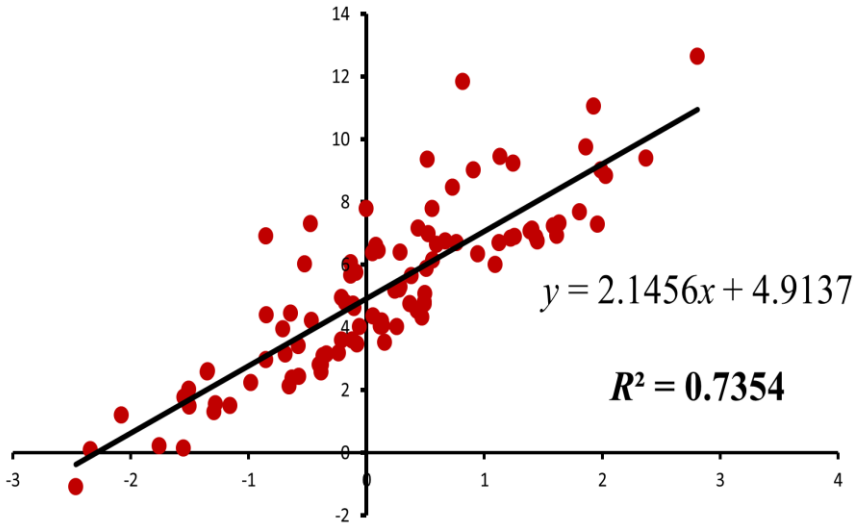
- Residual Plot:
  - 종속변수 Y에 대한 오차항(residual)은 설명변수 값의 범위에 관계없이 일정함(homoskedasticity)을 유지하는지 평가



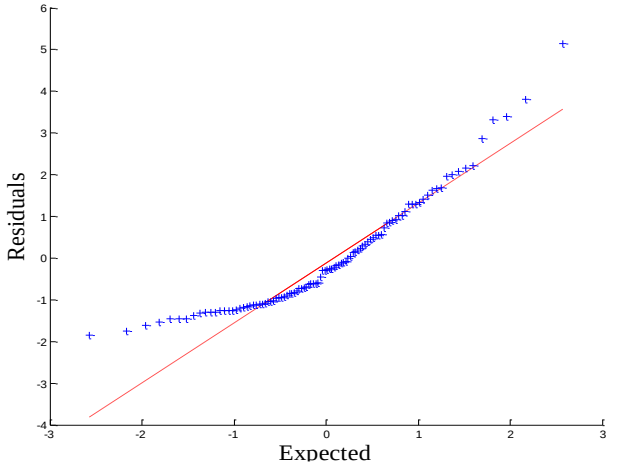
다중회귀분석: 모형의 적합도 평가

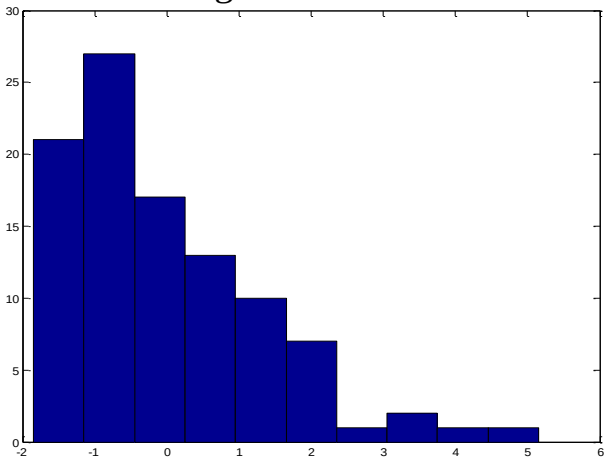
❖ 잔차의 정규성

$$y = 2x + \varepsilon, \qquad \varepsilon \sim Gamma(2,1)$$



Regression model





다중회귀분석: 예시

❖ 예시: 도요타 코롤라 중고차 가격 예측

Y

X

Price	Age_08_04	KM	Fuel_Type	HP	Met_Color	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185
19950	22	43610	Petrol	192	0	0	1800	3	100	1185
19600	25	32189	Petrol	192	0	0	1800	3	100	1185
21500	31	23000	Petrol	192	1	0	1800	3	100	1185
22500	32	34131	Petrol	192	1	0	1800	3	100	1185
22000	28	18739	Petrol	192	0	0	1800	3	100	1185
22750	30	34000	Petrol	192	1	0	1800	3	100	1185
17950	24	21716	Petrol	110	1	0	1600	3	85	1105
16750	24	25563	Petrol	110	0	0	1600	3	19	1065

다중회귀분석: 예시

❖ 데이터 전처리

- Fuel type 변수에 대한 1-of-C coding 변환

	Fuel_type = Disel	Fuel_type = Petrol	Fuel_type = CNG
Diesel	1	0	0
Petrol	0	1	0
CNG	0	0	1

❖ 데이터 구분

- 학습용 데이터 60%, 검증용 데이터 40%

Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type_Diesel	Fuel_Type_Petrol
1	RRA 2/3-Doors	13500	23	10	2002	46986	1	0
4	RRA 2/3-Doors	14950	26	7	2002	48000	1	0
5	SOL 2/3-Doors	13750	30	3	2002	38500	1	0
6	SOL 2/3-Doors	12950	32	1	2002	61000	1	0
9	VT I 2/3-Doors	21500	27	6	2002	19700	0	1
10	RRA 2/3-Doors	12950	23	10	2002	71138	1	0
12	BNS 2/3-Doors	19950	22	11	2002	43610	0	1
17	ORT 2/3-Doors	22750	30	3	2002	34000	0	1

다중회귀분석: 예시

❖ 다중회귀분석 결과물 해석

- 다중회귀분석을 수행하고 나면 다음과 같은 표를 결과로 얻을 수 있음

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

다중회귀분석: 예시

❖ 다중회귀분석 결과물 해석

- 회귀계수: Coefficient
  - ✓선형회귀분석에서 각 변수에 대응하는 베타값임
  - ✓해당 변수가 1단위 증가할 때 종속변수의 변화량을 의미
  - ✓양수이면 해당 설명변수와 종속변수는 양의 상관관계, 음수이면 음의 상관관계

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

다중회귀분석: 예시

❖ 다중회귀분석 결과물 해석

- 유의확률: p-value
  - ✓선형회귀분석에서 해당 변수가 통계적으로 유의미한지 알려주는 지표
  - ✓0에 가까울수록 모델링에 중요한 변수이며, 1에 가까울수록 유의미하지 않은 변수임
  - ✓특정 유의수준(α)을 설정하여 해당 값 미만의 변수만을 사용하여 다시 선형회귀분석을 구축하는 것도 가능함 (주로 $\alpha = 0.05$ 사용)

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

다중회귀분석: 예시

❖ 실제 종속변수 값과 예측된 종속변수 값의 차이(잔차) 분석

Predicted Value	Actual Value	Residual
15863.86944	13750	-2113.869439
16285.93045	13950	-2335.930454
16222.95248	16900	677.047525
16178.77221	18600	2421.227789
19276.03039	20950	1673.969611
19263.30349	19600	336.6965066
18630.46904	21500	2869.530964
18312.04498	22500	4187.955022
19126.94064	22000	2873.059357
16808.77828	16950	141.2217206
15885.80362	16950	1064.196384
15873.97887	16250	376.0211263
15601.22471	15750	148.7752903
15476.63164	15950	473.3683568
15544.83584	14950	-594.835836
15562.25552	14750	-812.2555172
15222.12869	16750	1527.871313
17782.33234	19000	1217.667664

목차

I 다중선형회귀분석

II 회귀분석 성능 평가

III 변수 선택

IV 다중선형회귀분석 예시

V R 실습

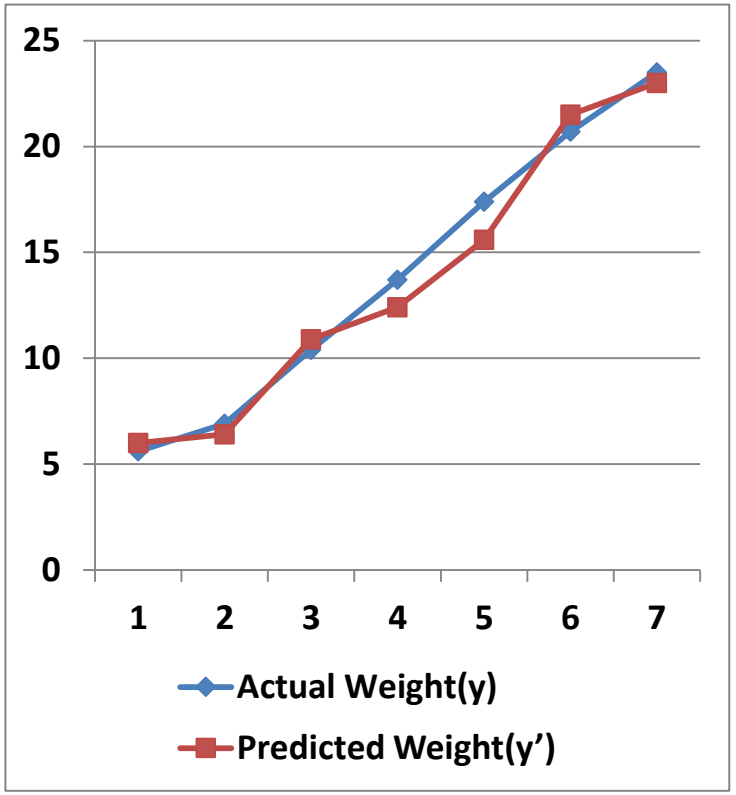
회귀분석 성능평가

I

예시

- 나이에 따른 아기의 몸무게 예측

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0



회귀분석 성능평가

2

성능지표 1:평균오차

- 실제 값에 비해 과대/과소 추정 여부를 판단
- 부호로 인해 잘못된 결론을 내릴 위험이 있음

$$Average\ error = \frac{1}{n}\sum_{i=1}^{n}(y - y') = 0.342$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

회귀분석 성능평가

3

성능지표 2: 평균 절대 오차(Mean absolute error; MAE)

- 실제 값과 예측 값 사이의 절대적인 오차의 평균을 이용

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - y'|$$
$$= 0.829$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

회귀분석 성능평가

4

성능지표 3: Mean absolute percentage error (MAPE)

- 실제값 대비 얼마나 예측 값이 차이가 있는지를 %로 표현
- 상대적인 오차를 추정하는데 주로 사용

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \frac{|y - y'|}{|y|}$$

$$= 6.43\%$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

회귀분석 성능평가

성능지표 4 & 5: (Root) Mean squared error ((R)MSE)

- 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 취한 지표

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y')^2$$
$$= 0.926$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y')^2}$$
$$= 0.962$$

Age	Actual Weight(y)	Predicted Weight(y')
1	5.6	6.0
2	6.9	6.4
3	10.4	10.9
4	13.7	12.4
5	17.4	15.6
6	20.7	21.5
7	23.5	23.0

5

회귀분석 성능평가

❖ 학습 및 검증 데이터에 대한 성능 평가

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

목차

I 다중선형회귀분석

II 회귀분석 성능 평가

III 변수 선택

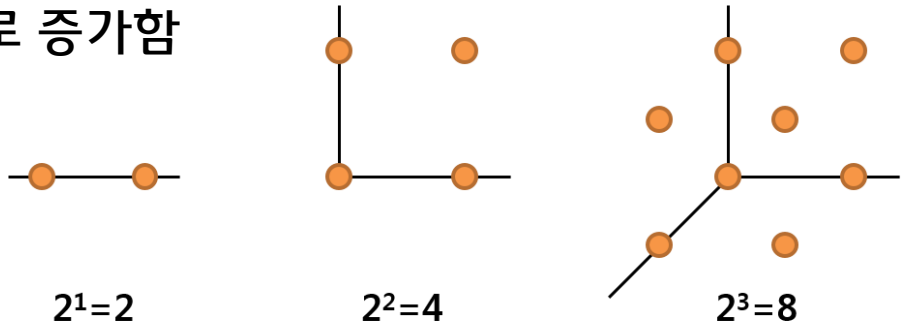
IV 다중선형회귀분석 예시

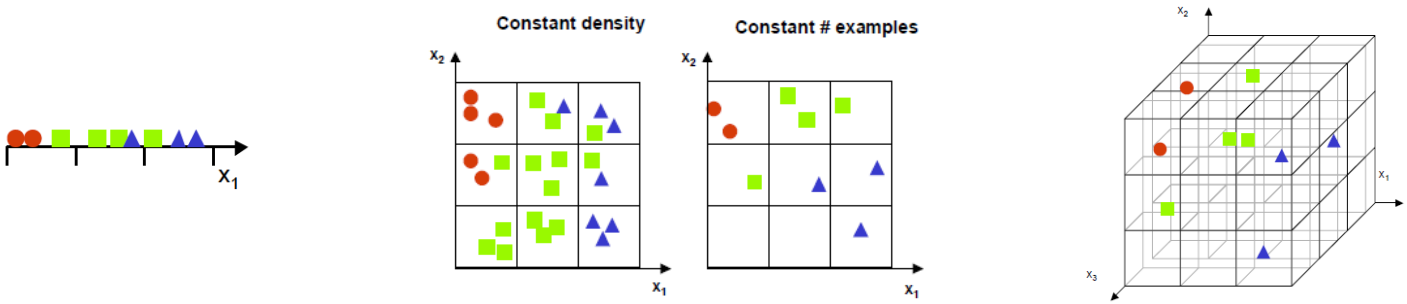
V R 실습

차원축소: Dimensionality Reduction

❖ 차원의 저주 (Curse of Dimensionality)

- 동등한 설명력을 갖기 위해서는 변수가 증가할 때 필요한 개체의 수는 기하급수적으로 증가함





"If there are various logical ways to explain a certain phenomenon, the simplest is the best" - Occam's Razor

차원축소: Dimensionality Reduction

❖ 차원축소: 배경

- 이론적으로는 변수의 수가 증가할수록 모델의 성능이 향상됨 (변수간 독립성 만족시)
- 실제 상황에서는 변수간 독립성 가정 위배, 노이즈 존재 등으로 인해 변수의 수가 일정 수준 이상 증가하면 모델의 성능이 저하되는 경향이 있음

❖ 차원축소: 목적

- 향후 분석 과정에서 성능을 저하시키지 않는 최소한의 변수 집합 판별

❖ 차원축소: 효과

- 변수간 상관성을 제거하여 결과의 통계적 유의성 제고
- 사후 처리(post-processing)의 단순화
- 주요 정보를 보존한 상태에서 중복되거나 불필요한 정보만 제거
- 고차원의 정보를 저차원으로 축소하여 시각화(visualization) 가능

차원축소: Dimensionality Reduction

❖ 차원축소 방식

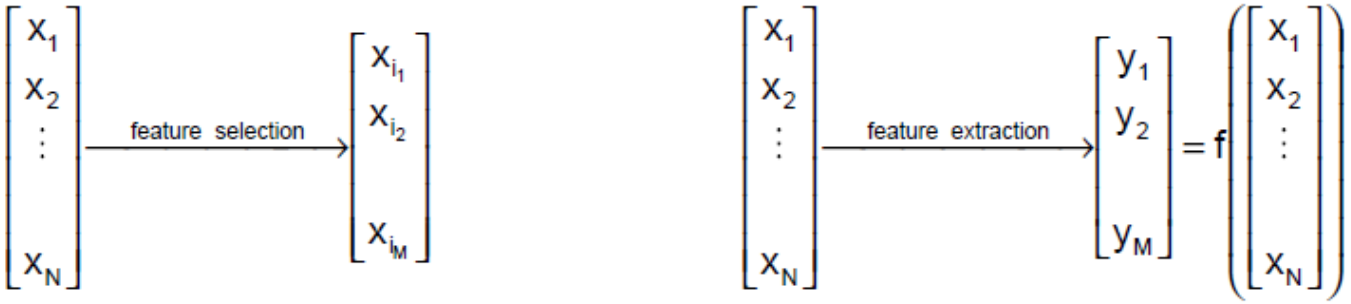
- 교사적 차원축소 (Supervised dimensionality reduction)
  - ✓축소된 차원의 적합성을 검증하는데 있어 데이터마이닝 모델을 적용
  - ✓동일한 데이터라도 적용되는 데이터마이닝 모델에 따라 축소된 차원의 결과가 달라질 수 있음

- 비교사적 차원축소 (Unsupervised dimensionality reduction)
  - ✓축소된 차원의 적합성을 검증하는데 있어 데이터마이닝 모델을 적용하지 않음
  - ✓특정 기법에 따른 차원축소 결과는 동일함

차원축소: Dimensionality Reduction

❖ 차원축소 기법

- 변수 선택(variable/feature selection)
  - ✓원래의 변수 집단으로부터 유용할 것으로 판단되는 소수의 변수들을 선택
  - ✓Filter – 변수 선택 과정과 모델 구축 과정이 독립적
  - ✓Wrapper – 변수 선택 과정이 데이터마이닝 모델의 결과를 최적화 하는 방향으로 이루어짐
- 변수 추출(variable/feature extraction)
  - ✓원래의 변수 집단을 보다 효율적인 적은 수의 새로운 변수 집단으로 변환
  - ✓데이터마이닝 모델에 독립적인 성능 지표가 추출된 변수의 효과를 측정하는 데 사용됨



차원축소: Dimensionality Reduction

❖ 차원 감소 기법 (cont')

- 변수 선택과 변수 추출 비교

X_1	X_2	X_3	...	X_n
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

변수 선택

X_1	X_5	X_8
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

변수 추출

Z_1	Z_2	Z_3
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

$Z_1 = X_1 + 0.2*X_2$

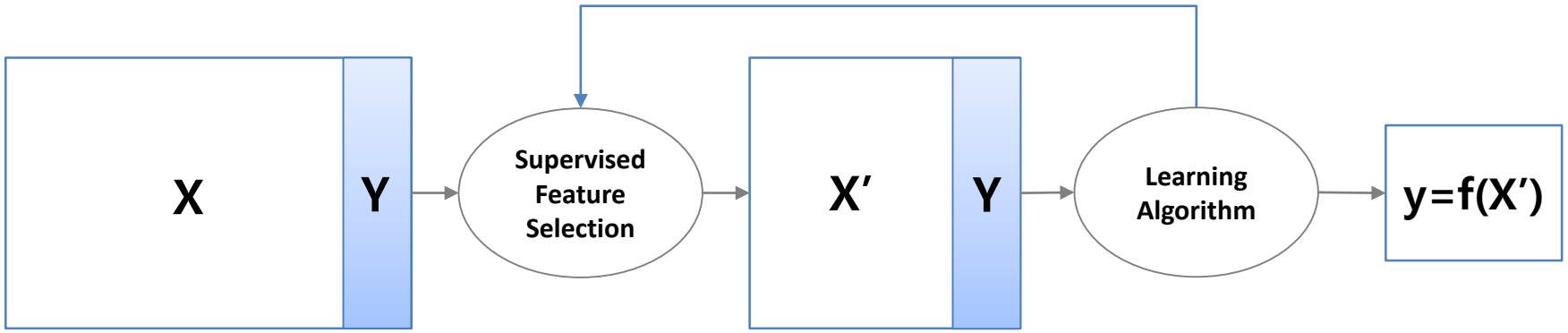
$Z_2 = X_3 - 2*X_5$

$Z_3 = X_4 + X_6 - X_9$

교사적 차원축소 기법

❖ 교사적 차원축소 기법 (Supervised feature selection)

- d-차원의 데이터에 대하여 사용하는 모델의 성능이 최대가 되도록 하는 d'차원(d' << d)의 변수를 선택

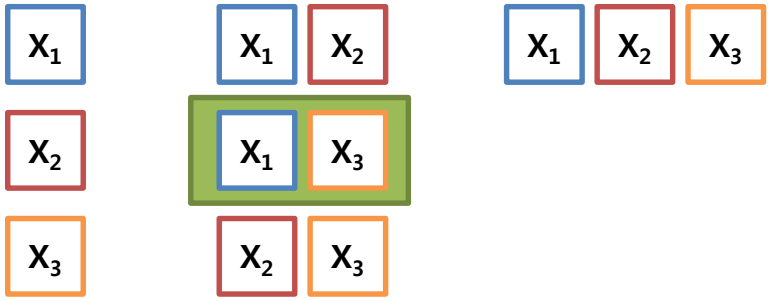


- 변수 선택을 하기 전, 모델 구축에 사용할 알고리즘을 먼저 선택
- 동일한 데이터라도 모델 구축에 사용되는 알고리즘에 따라 다양한 선택 결과가 나타날 수 있음

Forward/Backward/Stepwise Selection

❖ 전역 탐색 (Exhaustive search)

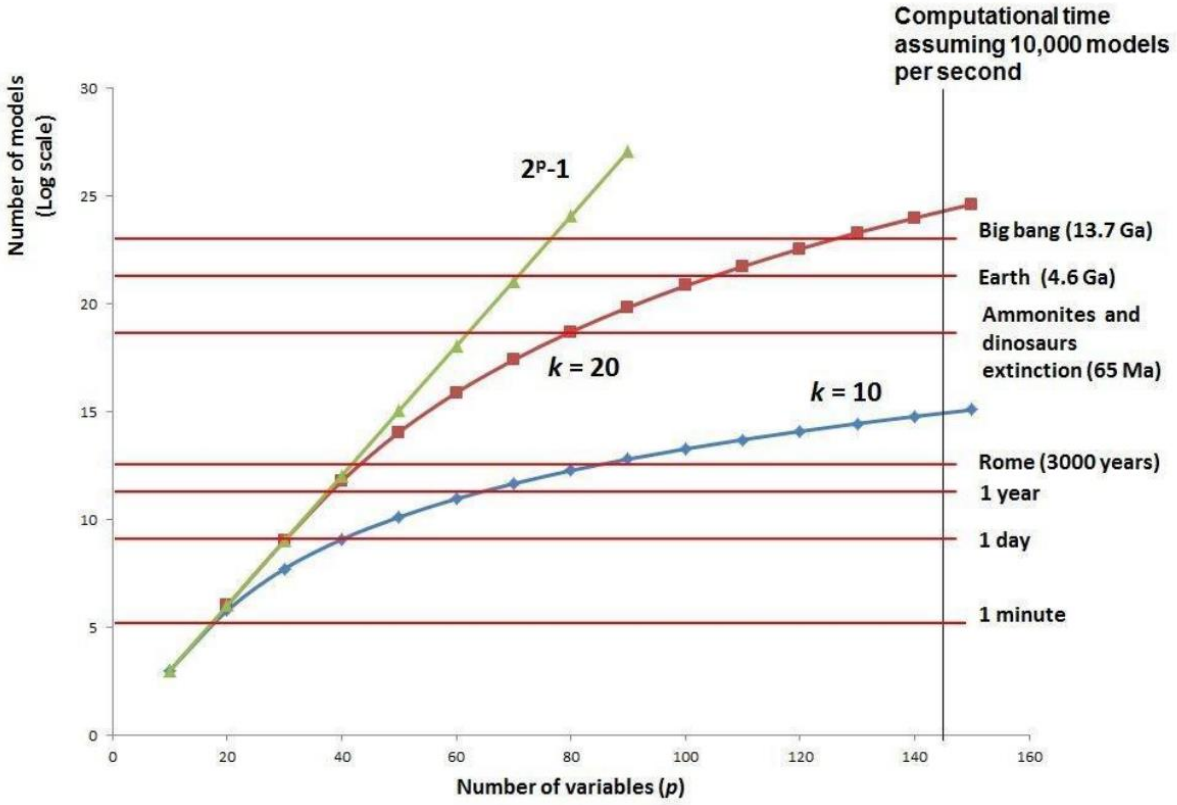
- 가능한 모든 경우의 조합에 대해 모델을 구축한 뒤 최적의 변수 조합을 찾는 방식
  - ✓ 예: 3개의 변수가 존재하는 경우 X_1 X_2 X_3
  - ✓ 총 여섯 가지의 가능한 변수 조합 존재



- 변수 선택을 위한 모델 평가 기준
  - ✓ Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), 수정 R-제곱합, Mallow's C_p 등

Forward/Backward/Stepwise Selection

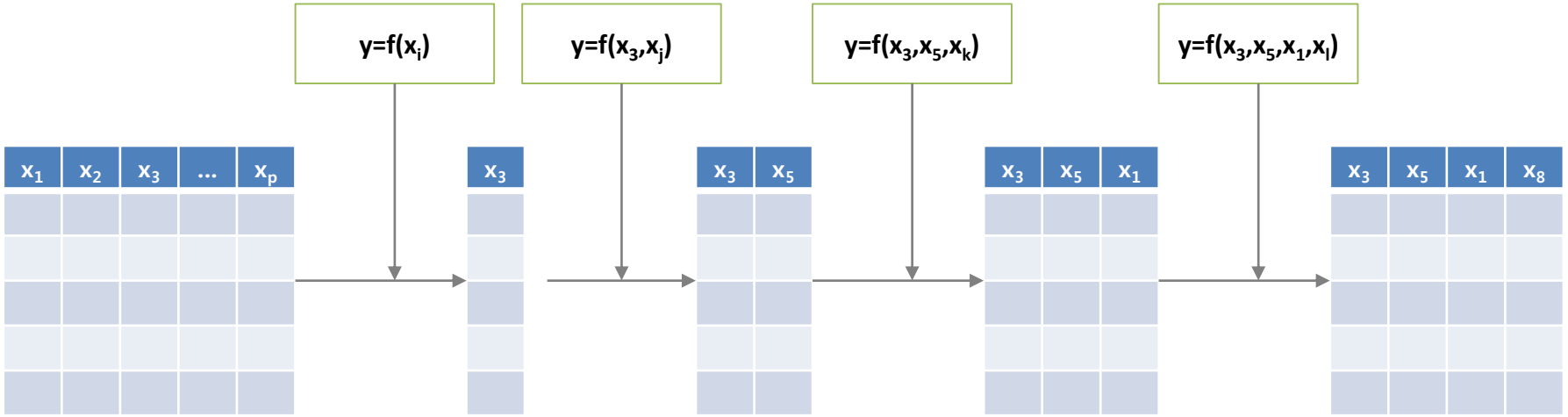
- 전역 탐색 (Exhaustive search)
  - 1초에 10,000개의 모델을 평가할 수 있는 컴퓨터를 활용할 경우 변수 선택에 소요되는 시간



전진 선택법 (Forward Selection)

❖ 전진 선택법

- 설명변수가 하나도 없는 모델에서부터 시작하여 가장 유의미한 변수를 하나씩 추가해 나가는 방법 (회귀분석 모델의 F-통계량 사용)
- 한번 선택된 변수는 제거되지 않음 (변수의 숫자는 단조 증가)
- 전진 선택법 예시



전진 선택법 (Forward Selection)

❖ 전진 선택법

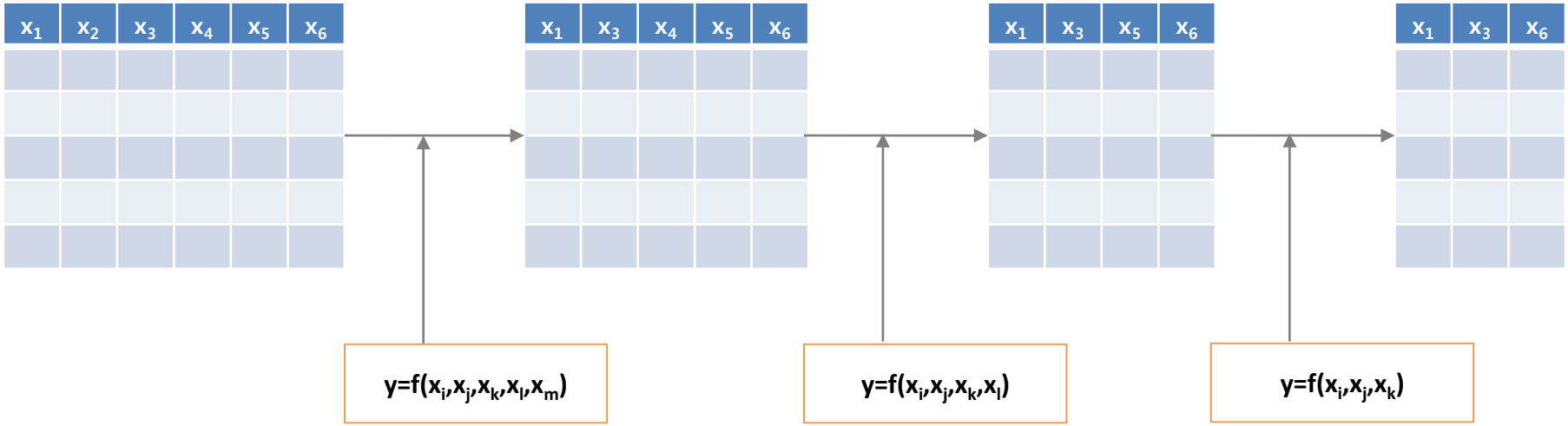
- 선형회귀분석에서의 전진선택법

```
Step 5 Var CHEST Entered R-sq=0.5379 C(p)= 4.195
             DF   Sum Sq  Mean Sq  F  Prob>F
Regression    5  108.3272  21.6654  24.91 0.0001
Error       107   93.0527   0.86965099
Total       112  201.37982301
           Par    Std      Type II
Variable    Est  Error   Sum Sq     F   Prob>F
INTERCEP -0.7680 0.6102   1.3776   1.58 0.2109
CULTURE   0.0432 0.0098  16.7198  19.23 0.0001
STAY      0.2339 0.0574  14.4381  16.60 0.0001
NRATIO    0.6724 0.2993   4.3888   5.05 0.0267
CHEST     0.0092 0.0054   2.5062   2.88 0.0925
FACIL     0.0184 0.0063   7.4571   8.57 0.0042
```

후진 소거법 (Backward Elimination)

❖ 후진 소거법

- 모든 변수를 사용하여 구축한 모델에서 유의미하지 않은 변수를 하나씩 제거해 나가는 방법
- 한번 제거된 변수는 다시 선택될 가능성이 없음 (변수의 숫자는 단조 증가)
- 후진 소거법 예시



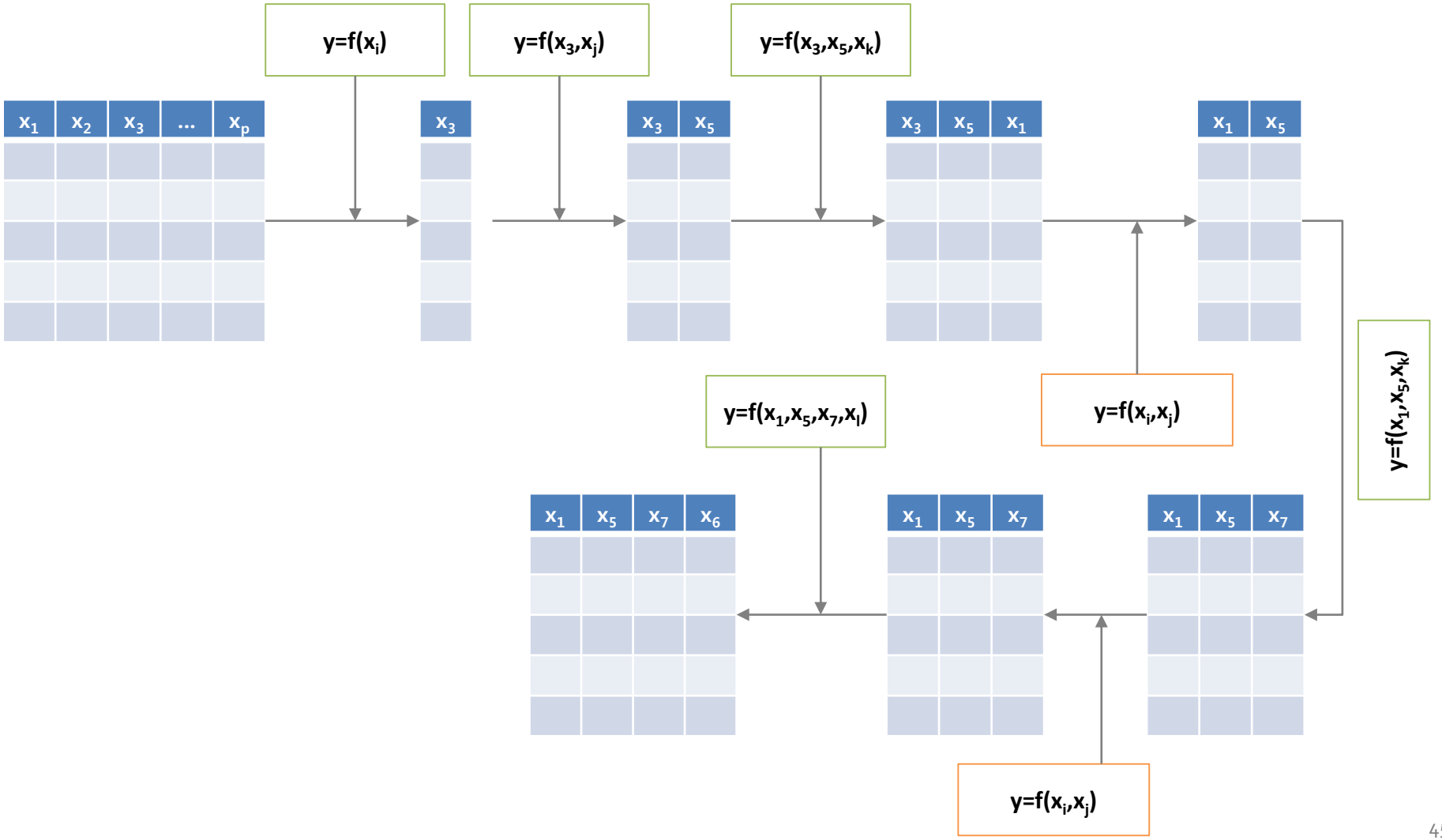
단계적 선택법 (Stepwise Selection)

❖ 단계적 선택법

- 설명변수가 하나도 없는 모델에서부터 시작하여 전진선택법과 후진소거법을 번갈아가며 수행
- 전진선택법 및 후진소거법에 비해 시간을 오래 걸리나 보다 우수한 예측 성능을 나타내는 변수 집합을 찾아낼 가능성이 높음
- 한번 선택되거나 제거된 변수라도 다시 선택/제거될 가능성이 있음
- 변수의 수는 초기에는 일반적으로 증가하나 중반 이후에는 증가와 감소를 반복

단계적 선택법 (Stepwise Selection)

❖ 단계적 선택법 예시



변수선택 평가지표

❖ 변수선택 평가지표 1 & 2

- Akaike Information Criteria (AIC)
  - ✓잔차제곱합에 변수의 수를 penalty term으로 추가

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2k$$

- Bayesian Information Criteria (BIC)
  - ✓잔차제곱합, 사용 변수의 수, 모든변수를 사용한 모델에서 추정된 잔차의 표준편차를

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2}$$

변수선택 평가지표

❖ 변수선택 평가지표 3

- 수정 R-제곱합 (Adjusted R^2)
  - ✓ 단순 R-제곱합은 변수가 많아질수록 증가하므로 변수선택에 사용하기 좋은 평가지표가 아님

Model 1 : $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$

Model 2 : $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \ldots + \beta_{k+m} x_{k+m} \epsilon$

$$R^2(M2) \geq R^2(M1)$$

  - ✓ 변수의 수(k)를 고려한 수정 R-제곱합을 사용

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2) = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSE}{SSTot}$$

변수선택 평가지표

❖ 변수선택 평가지표 4

- Mallow's C_p

✓ 모델에 의해 설명되지 못하는 오차는 편기(Bias)와 분산(Variance)로 분해할 수 있음

$$\hat{y}_i - \mu_i = (E[\hat{y}_i] - \mu_i) + (\hat{y}_i - E[\hat{y}_i])$$
$$= \text{Bias} + \text{Random error}$$

$$E[(\hat{y}_i - \mu_i)^2] = (E[\hat{y}_i] - \mu_i)^2 + \text{Var}(\hat{y}_i)$$
$$= \text{Bias}^2 + \text{Variance}$$

✓ 다음과 같은 유도 과정을 통해 Mallow's C_p를 최소화하는 변수 집합이 모델의 예측 성능을 최대화하는 것을 알 수 있음

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(E[\hat{y}_i] - \mu_i)^2 + \sum_{i=1}^{n}\text{Var}(\hat{y}_i)\right\}$$
$$= \frac{SSB(p)}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Var}(\hat{y}_i)$$

$$\Gamma_p = \frac{1}{\sigma^2}\left\{E[SSE(p)] - (n-p)\sigma^2 + p\sigma^2\right\}$$
$$= \frac{E[SSE(p)]}{\sigma^2} - n + 2p$$

p개의 선택된 변수를 이용한 모델의 SSE

$$C_p = \frac{SSE(p)}{MSE(K+1)} - n + 2p$$

모든 변수를 이용한 모델의 MSE

선형 회귀분석에서의 변수선택

❖ 전역 탐색 결과

#Coeffs	RSS	Cp	R^2	Adj. R^2	Prob	Model (Constant present in all models)											
						1	2	3	4	5	6	7	8	9	10	11	12
2	1,996,467,712	477.712	0.747	0.746	0.000	Constant	Age	*	*	*	*	*	*	*	*	*	*
3	1,672,546,432	305.506	0.788	0.787	0.000	Constant	Age	HP	*	*	*	*	*	*	*	*	*
4	1,438,242,432	181.495	0.818	0.817	0.000	Constant	Age	HP	Weight	*	*	*	*	*	*	*	*
5	1,258,062,976	86.594	0.840	0.839	0.000	Constant	Age	Mileage	HP	Weight	*	*	*	*	*	*	*
6	1,181,816,320	47.588	0.850	0.849	0.000	Constant	Age	Mileage	Petrol	Quarterly_Tax	Weight	*	*	*	*	*	*
7	1,095,153,024	2.980	0.861	0.860	0.962	Constant	Age	Mileage	Petrol	HP	Quarterly_Tax	Weight	*	*	*	*	*
8	1,093,753,344	4.227	0.861	0.860	0.994	Constant	Age	Mileage	Petrol	HP	Automatic	Quarterly_Tax	Weight	*	*	*	*
9	1,093,557,120	6.122	0.861	0.859	0.989	Constant	Age	Mileage	Petrol	HP	Metalic_Color	Automatic	Quarterly_Tax	Weight	*	*	*
10	1,093,422,592	8.049	0.861	0.859	0.976	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	Quarterly_Tax	Weight	*	*
11	1,093,335,424	10.002	0.861	0.859	0.961	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	CC	Quarterly_Tax	Weight	*
12	1,093,331,072	12.000	0.861	0.859	1.000	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	CC	Doors	Quarterly_Tax	Weight

선형 회귀분석에서의 변수선택

❖ 후진소거법 결과

#Coeffs	RSS	Cp	R^2	Adj. R^2	Prob	Model (Constant present in all models)											
						1	2	3	4	5	6	7	8	9	10	11	12
2	1,996,467,712	477.712	0.747	0.746	0.000	Constant	Age	*	*	*	*	*	*	*	*	*	*
3	1,780,184,064	363.394	0.774	0.773	0.000	Constant	Age	Weight	*	*	*	*	*	*	*	*	*
4	1,482,806,272	205.462	0.812	0.811	0.000	Constant	Age	Petrol	Weight	*	*	*	*	*	*	*	*
5	1,310,214,400	114.641	0.834	0.833	0.000	Constant	Age	Petrol	terly_Tax	Weight	*	*	*	*	*	*	*
6	1,181,816,320	47.588	0.850	0.849	0.000	Constant	Age	Mileage	Petrol	Quarterly_Tax	Weight	*	*	*	*	*	*
7	1,095,153,024	2.980	0.861	0.860	0.962	Constant	Age	Mileage	Petrol	HP	Quarterly_Tax	Weight	*	*	*	*	*
8	1,093,753,344	4.227	0.861	0.860	0.994	Constant	Age	Mileage	Petrol	HP	Automatic	Quarterly_Tax	Weight	*	*	*	*
9	1,093,557,120	6.122	0.861	0.859	0.989	Constant	Age	Mileage	Petrol	HP	Metalic_Color	Automatic	Quarterly_Tax	Weight	*	*	*
10	1,093,422,592	8.049	0.861	0.859	0.976	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	Quarterly_Tax	Weight	*	*
11	1,093,335,424	10.002	0.861	0.859	0.961	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	CC	Quarterly_Tax	Weight	*
12	1,093,331,072	12.000	0.861	0.859	1.000	Constant	Age	Mileage	Diesel	Petrol	HP	Metalic_Color	Automatic	CC	Doors	Quarterly_Tax	Weight

선형 회귀분석에서의 변수선택

❖ 선택된 변수를 이용한 회귀분석 모델

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3874.492188	1415.003052	0.00640071	97276411904
Age_08_04	-123.4366303	3.33806777	0	8033339392
KM	-0.01749926	0.00173714	0	251574528
Fuel_Type_Petrol	2409.154297	319.5795288	0	5049567
HP	19.70204735	4.22180223	0.00000394	291336576
Quarterly_Tax	16.88731384	2.08484554	0	192390864
Weight	15.91809368	1.26474357	0	281026176

Model Fit

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1516825972	1326.521353	-0.000143957

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

Predictive performance

(compare to 12-predictor model!)

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1021510219	1334.029433	118.4483556

Validation Data scoring - Summary Report

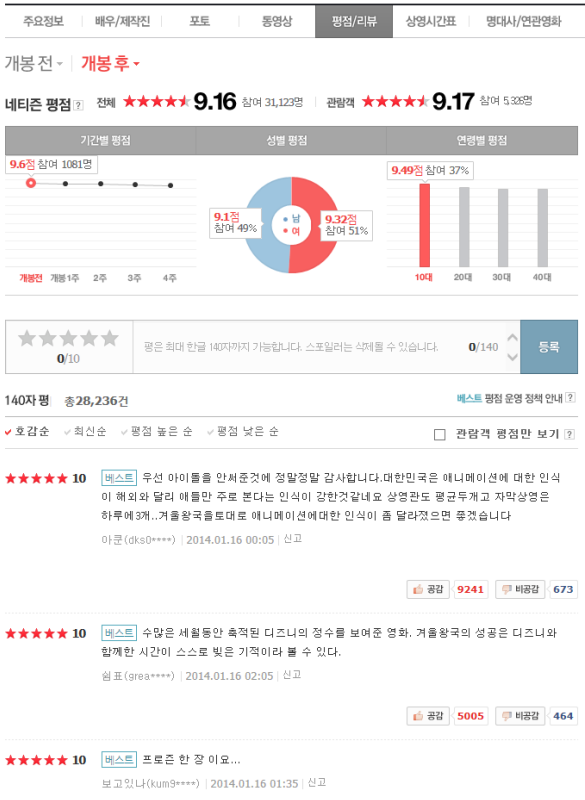
Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

목차

I 다중선형회귀분석
II 회귀분석 성능 평가
III 변수 선택
IV 다중선형회귀분석 예시
V R 실습

다중선형회귀분석 예시

❖ SNS 검색어를 활용한 영화 총수입 예측



다중선형회귀분석 예시

❖ SNS 검색어를 활용한 영화 총수입 예측

가정:
개봉 전과 개봉 초기에 SNS에 올려진 mention들은 영화의 기초정보(상영 스크린 수, 전주 box office 수입 등)와 결합하여 영화의 총 수입을 일정 수준 이상으로 정확하게 예측할 수 있을 것이다!

다중선형회귀분석 예시

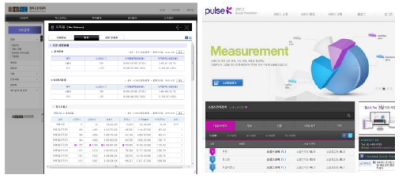
❖ SNS 검색어를 활용한 영화 총수입 예측

1 Movie Selection & Data Collection

Filtering rules
- Not too general title (ex: Mother)
- Total number of audiences ≥ 100,000

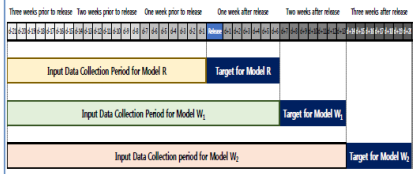
Screening data
- KOFIC
- www.kobis.or.kr

SNS data
- pulseK
- www.pulsek.com



2 Model Configuration & Variable Definition

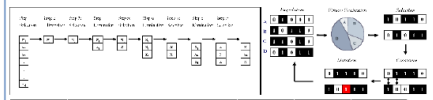
Model configuration
- Three forecasting models at different forecasting time and data collection periods



Variable definition
- Model **R** : 25 variables (1 screening, 24 SNS)
- Model W_1: 32 variables (4 screening, 28 SNS)
- Model W_2: 38 variables (6 screening, 32 SNS)

3 Variable Selection

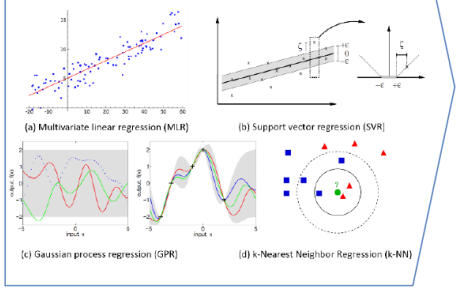
Variable selection
- Perform stepwise selection for forecasting models with screening data only
- Perform genetic algorithm for forecasting models with screening & SNS data



		w/o SNS	with SNS			
		Baseline	MLR	SVR	GPR	k-NN
Model R	Screening	1	1	1	1	1
	SNS		7	7	3	5
Model W_1	Screening	2	2	2	4	2
	SNS		2	6	0	8
Model W_2	Screening	4	3	6	5	6
	SNS		7	6	10	9

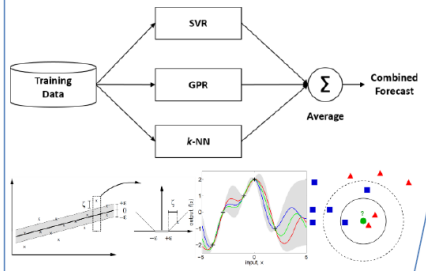
4 Single Forecasting Model Development & Evaluation

Single forecasting models
- 1 linear algorithm
- 3 machine learning-based algorithms



5 Combining Forecasts & Evaluation

Forecasting Combination
- Individual models: machine learning-based algorithms
- Combining rule: Equally weighted average



다중선형회귀분석 예시

❖ SNS 검색어를 활용한 영화 총수입 예측

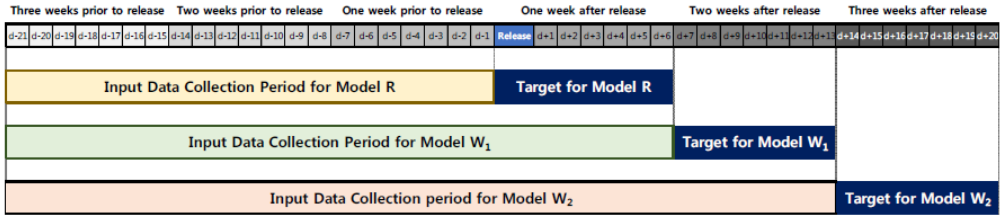


Figure 2: The timing of forecasting and data collection period of each forecasting model.

Table 4: The candidate input variables for each forecasting model.

Category	Attribute	Model R	Model W_1	Model W_2
Target		$\text{Log}_{10}(\text{Box\_office}^1)$	$\text{Log}_{10}(\sum_{i=1}^{2}\text{Box\_office}^i)$	$\text{Log}_{10}(\sum_{i=1}^{3}\text{Box\_office}^i)$
Screening	Original	N_seat_1	Box_office^1 N_seat_2 N_seat^1	Box_office^1+Box_office^2 N_seat_3 N_seat^2
	Derived		Weekly_seat_increase	N_seat^1+N_seat^2 Weekly_seat_increase Weekly_cumulative_seat_increase
SNS	Original	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3,-2,-1\}$	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3,-2,-1,1\}$	N_mention^p N_emotional^p N_positive^p N_negative^p $p \in \{-3,-2,-1,1,2\}$
	Derived	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase	Cumulative_mention Cumulative_emotional Cumulative_positive Cumulative_negative Avg_mention_increase Weekly_mention_increase Avg_emotional_increase Weekly_emotional_increase Avg_positive_increase Weekly_positive_increase Avg_negative_increase Weekly_negative_increase

다중선형회귀분석 예시

❖ SNS 검색어를 활용한 영화 총수입 예측

Model	Linear Algorithm		Machine Learning				Combination
	Baseline	MLR	SVR	GPR	k-NN	Average	
Model R	0.6219	0.4366 (29.80%)	0.4015 (35.45%)** (8.05%)$^{+}$	0.4093 (34.19%)*** (6.25%)$^{++}$	0.3984 (35.93%)* (8.74%)$^{+}$	0.4031 (35.19%) (7.68%)	0.3933 (36.76%)*** (9.92%)$^{++}$ (2.42%)
Model W_1	0.1541	0.1160 (24.74%)	0.1084 (29.65%)* (6.53%)	0.1087 (29.46%)* (6.27%)	0.1154 (25.16%) (0.56%)	0.1108 (28.09%) (4.45%)	0.1058 (31.36%)** (8.79%) (4.54%)
Model W_2	0.0722	0.0533 (26.14%)	0.0425 (41.16%) (20.33%)	0.0440 (39.07%)*** (17.50%)$^{+++}$	0.0496 (31.26%) (6.93%)	0.0453 (37.16%) (14.92%)	0.0419 (41.94%)*** (21.39%)$^{+++}$ (7.60%)

➔ MAPE를 성능평가지표로 사용했을 때, 선형회귀분석(MLR)은 개봉 시점에서 총수입을 43% 오차 이내로 예측할 수 있으며, 개봉 2주 후에는 그 오차가 5% 내외로 감소함

다중선형회귀분석

❖ 요약

- 다중선형회귀분석은 탐색적 목적 뿐만 아니라 예측적 목적 측면에서도 매우 유용하게 사용되는 방법론임
- 예측 모델은 학습 데이터를 이용하여 회귀계수를 추정하고 이를 학습에 사용하지 않은 검증 데이터에 적용하여 그 성능을 평가함
- 불필요하거나 중복된 설명변수를 적절히 제거하는 것이 우수한 예측 정확도와 회귀 모델의 강건성(robustness)을 확보하는데 매우 중요함
- 변수선택 기법을 통해 모델의 예측 성능을 저하시키지 않는 효율적이면서 적은 수의 설명변수들의 집합을 찾아낼 수 있음

목차

I 다중선형회귀분석

II 회귀분석 성능 평가

III 변수 선택

IV 다중선형회귀분석 예시

V R 실습

R 실습: 다중선형회귀분석 및 변수선택

❖ 선형회귀분석 데이터: Toyota Corolla 중고차가격 예측

Variable	Description	Variable	Description
		Guarantee_Period	Guarantee period in months
		ABS	Anti-Lock Brake System (Yes=1, No=0)
Price	Offer Price in EUROs	Airbag_1	Driver_Airbag (Yes=1, No=0)
Age_08_04	Age in months as in August 2004	Airbag_2	Passenger Airbag (Yes=1, No=0)
Mfg_Month	Manufacturing month (1-12)	Airco	Airconditioning (Yes=1, No=0)
Mfg_Year	Manufacturing Year	Automatic_airco	Automatic Airconditioning (Yes=1, No=0)
KM	Accumulated Kilometers on odometer	Boardcomputer	Boardcomputer (Yes=1, No=0)
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)	CD_Player	CD Player (Yes=1, No=0)
HP	Horse Power	Central_Lock	Central Lock (Yes=1, No=0)
Met_Color	Metallic Color? (Yes=1, No=0)	Powered_Windows	Powered Windows (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)	Power_Steering	Power Steering (Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters	Radio	Radio (Yes=1, No=0)
Doors	Number of doors	Mistlamps	Mistlamps (Yes=1, No=0)
Cylinders	Number of cylinders	Sport_Model	Sport Model (Yes=1, No=0)
Gears	Number of gear positions	Backseat_Divider	Backseat Divider (Yes=1, No=0)
Quarterly_Tax	Quarterly road tax in EUROs	Metallic_Rim	Metallic Rim (Yes=1, No=0)
Weight	Weight in Kilograms	Radio_cassette	Radio Cassette (Yes=1, No=0)
Mfr_Guarantee	Within Manufacturer's Guarantee period (Yes=1, No=0)	Parking_Assistant	Parking assistance system (Yes=1, No=0)
BOVAG_Guarantee	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)	Tow_Bar	Tow Bar (Yes=1, No=0)

R 실습: 다중선형회귀분석 및 변수선택

❖ 데이터 불러오기 & 전처리

- 범주형 변수(선형회귀분석 사용 불가능)를 이진형 변수로 변환

```r
# Working directory 지정
setwd("C:\\RStudy")

# 실습 1: 전진선택/후진소거/단계적선택 ----------------------
# 분석에 필요한 패키지 설치 및 불러오기
# Multivariate linear regression
corolla <- read.csv("ToyotaCorolla.csv")

# Indices for the activated input variables
nCar <- dim(corolla)[1]
nVar <- dim(corolla)[2]

id_idx <- c(1,2)
category_idx <- 8

# 범주형 변수를 이진형 변수로 변환
dummy_p <- rep(0,nCar)
dummy_d <- rep(0,nCar)
dummy_c <- rep(0,nCar)

p_idx <- which(corolla$Fuel_Type == "Petrol")
d_idx <- which(corolla$Fuel_Type == "Diesel")
c_idx <- which(corolla$Fuel_Type == "CNG")

dummy_p[p_idx] <- 1
dummy_d[d_idx] <- 1
dummy_c[c_idx] <- 1

Fuel <- data.frame(dummy_p, dummy_d, dummy_c)
names(Fuel) <- c("Petrol","Diesel","CNG")

# Prepare the data for MLR
mlr_data <- cbind(corolla[,-c(id_idx, category_idx)], Fuel)
```

Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type	HP	Met_Color	Automatic	cc
13500	23	10	2002	46986	Diesel	90	1	0	2000
13750	23	10	2002	72937	Diesel	90	1	0	2000
13950	24	9	2002	41711	Diesel	90	1	0	2000
14950	26	7	2002	48000	Diesel	90	0	0	2000
13750	30	3	2002	38500	Diesel	90	0	0	2000
12950	32	1	2002	61000	Diesel	90	0	0	2000
16900	27	6	2002	94612	Diesel	90	1	0	2000
18600	30	3	2002	75889	Diesel	90	1	0	2000
21500	27	6	2002	19700	Petrol	192	0	0	1800
12950	23	10	2002	71138	Diesel	69	0	0	1900
20950	25	8	2002	31461	Petrol	192	0	0	1800
19950	22	11	2002	43610	Petrol	192	0	0	1800
19600	25	8	2002	32189	Petrol	192	0	0	1800
21500	31	2	2002	23000	Petrol	192	1	0	1800
22500	32	1	2002	34131	Petrol	192	1	0	1800

KM	HP	Met_Color
46986	90	1
72937	90	1
41711	90	1
48000	90	0
38500	90	0
61000	90	0
94612	90	1
75889	90	1
19700	192	0
71138	69	0
31461	192	0
43610	192	0
32189	192	0
23000	192	1

...

Petrol	Diesel	CNG
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
1	0	0
0	1	0
1	0	0
1	0	0
1	0	0
1	0	0

R 실습: 다중선형회귀분석 및 변수선택

❖ 데이터를 학습용/검증용으로 분할

```
35  # Split the data into the training/validation sets
36  trn_idx <- sample(1:nCar, round(0.7*nCar))
37  trn_data <- mlr_data[trn_idx,]
38  val_data <- mlr_data[-trn_idx,]
```

mlr_data	1436 obs. of 37 variables
trn_data	1005 obs. of 37 variables
val_data	431 obs. of 37 variables

❖ 모든 변수를 사용한 선형회귀분석 모델 구축

```
40  # Train the MLR
41  full_model <- lm(Price ~ ., data = trn_data)
42  full_model
43  summary(full_model)
44  plot(full_model)
45
46  # Plot the result
47  plot(trn_data$Price, fitted(full_model), xlim = c(4000,35000), ylim = c(4000,35000))
48  abline(0,1,lty=3)
49
50  anova(full_model)
51  plot(fitted(full_model), resid(full_model), xlab="Fitted values", ylab="Residuals")
```

R 실습: 다중선형회귀분석 및 변수선택

❖ 모든 데이터를 이용한 선형회귀분석 결과

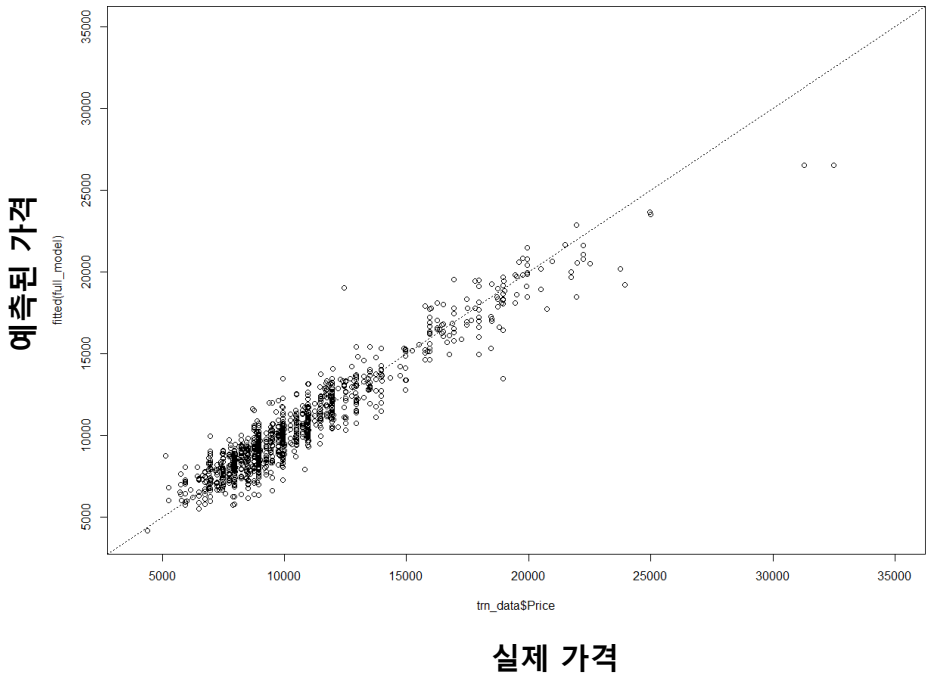
```
> summary(full_model)

Call:
lm(formula = Price ~ ., data = trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6571.9  -640.9   -49.0   624.2  5972.3

Coefficients: (3 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.540e+03  1.724e+03   2.054 0.040257 *
Age_08_04        -1.178e+02  3.914e+00 -30.104  < 2e-16 ***
Mfg_Month        -1.059e+02  1.034e+01 -10.244  < 2e-16 ***
Mfg_Year                NA         NA      NA       NA
KM               -1.710e-02  1.338e-03 -12.777  < 2e-16 ***
HP                1.911e+01  3.601e+00   5.305 1.39e-07 ***
Met_Color        -4.358e+01  7.632e+01  -0.571 0.568134
Automatic         3.746e+02  1.458e+02   2.568 0.010368 *
cc               -5.613e-02  7.515e-02  -0.747 0.455279
Doors             7.198e+01  4.111e+01   1.751 0.080257 .
Cylinders               NA         NA      NA       NA
Gears             1.959e+02  2.142e+02   0.915 0.360617
Quarterly_Tax     1.159e+01  2.128e+00   5.446 6.52e-08 ***
Weight            8.879e+00  1.227e+00   7.233 9.54e-13 ***
Mfr_Guarantee     2.360e+02  7.381e+01   3.198 0.001430 **
BOVAG_Guarantee   3.989e+02  1.316e+02   3.033 0.002490 **
Guarantee_Period  7.207e+01  1.459e+01   4.938 9.27e-07 ***
ABS              -4.715e+01  1.300e+02  -0.363 0.716844
Airbag_1          4.498e+02  2.570e+02   1.750 0.080375 .
Airbag_2         -2.007e+02  1.314e+02  -1.527 0.127121
Airco             2.245e+02  8.919e+01   2.517 0.012008 *
Automatic_airco   2.435e+03  1.889e+02  12.890  < 2e-16 ***
Boardcomputer    -2.099e+02  1.194e+02  -1.758 0.078992 .
CD_Player         8.442e+01  1.010e+02   0.836 0.403239
Central_Lock     -7.471e+01  1.419e+02  -0.526 0.598678
Powered_Windows   5.112e+02  1.424e+02   3.589 0.000349 ***
Power_Steering   -5.689e+02  2.842e+02  -2.002 0.045581 *
Radio             5.575e+02  6.295e+02   0.886 0.376037
Mistlamps         1.869e+01  1.102e+02   0.170 0.865286
Sport_Model       2.790e+02  8.906e+01   3.132 0.001787 **
Backseat_Divider -6.961e+01  1.327e+02  -0.525 0.599953
Metallic_Rim      5.536e+01  9.675e+01   0.572 0.567342
Radio_cassette   -5.593e+02  6.299e+02  -0.888 0.374863
Tow_Bar          -1.990e+02  8.018e+01  -2.482 0.013216 *
Petrol            1.096e+03  4.339e+02   2.527 0.011663 *
Diesel            5.269e+02  4.128e+02   1.276 0.202180
CNG                     NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1060 on 971 degrees of freedom
Multiple R-squared:  0.9127, Adjusted R-squared:  0.9098
F-statistic: 307.7 on 33 and 971 DF,  p-value: < 2.2e-16
```



R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 I: 전진선택법

- 상수항만 존재하는 모델에서 중요한 변수를 하나씩 추가

```
53 # 변수선택 1: 전진선택법
54 # Upperbound formula 만들기
55 tmp_x <- paste(colnames(trn_data)[-1], collapse=" + ")
56 tmp_xy <- paste("Price ~ ", tmp_x, collapse = "")
57 tmp_xy
58 as.formula(tmp_xy)
59
60 forward_model <- step(lm(Price ~ 1, data = trn_data),
61                       scope = list(upper = as.formula(tmp_xy), lower = Price ~ 1), direction="forward", trace=1)
62 summary(forward_model)
63 anova(forward_model)
64
65 # 각 단계에서 선택된 변수 표시
66 forward_model$anova$Step
67 forward_model$anova$AIC
68
69 # 선택된 변수에 따른 AIC 감소분 표시
70 plot(forward_model$anova$AIC, pch = 17, cex=2, main = "AIC Decrease (Forward Selection)", xlab = "Number of Steps", ylab = "AIC")
71 text(forward_model$anova$AIC, forward_model$anova$Step, cex=1, pos=3, col="blue")
```

R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 I: 전진선택법

- 변수선택 결과 (최초 36개 변수 → 20개 변수 선택)

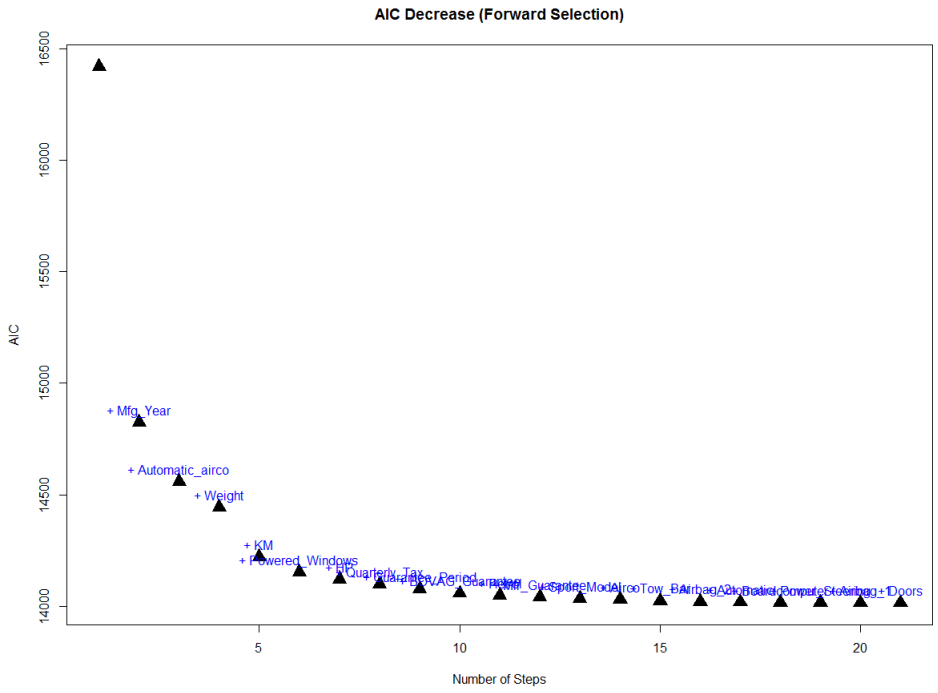
```
> summary(forward_model)

Call:
lm(formula = Price ~ Mfg_Year + Automatic_airco + Weight + KM +
    Powered_Windows + HP + Quarterly_Tax + Guarantee_Period +
    BOVAG_Guarantee + Petrol + Mfr_Guarantee + Sport_Model +
    Airco + Tow_Bar + Airbag_2 + Automatic + Boardcomputer +
    Power_Steering + Airbag_1 + Doors, data = trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6747.2  -653.8   -53.8   640.8  5908.7

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.807e+06  8.542e+04 -32.857  < 2e-16 ***
Mfg_Year         1.402e+03  4.286e+01  32.718  < 2e-16 ***
Automatic_airco  2.451e+03  1.746e+02  14.037  < 2e-16 ***
Weight           9.233e+00  1.166e+00   7.918 6.50e-15 ***
KM              -1.734e-02  1.309e-03 -13.252  < 2e-16 ***
Powered_Windows  4.650e+02  8.300e+01   5.602 2.74e-08 ***
HP               1.819e+01  3.235e+00   5.625 2.42e-08 ***
Quarterly_Tax    1.146e+01  2.013e+00   5.694 1.63e-08 ***
Guarantee_Period 7.624e+01  1.377e+01   5.535 3.98e-08 ***
BOVAG_Guarantee  4.078e+02  1.268e+02   3.216 0.001342 **
Petrol           6.593e+02  2.956e+02   2.231 0.025933 *
Mfr_Guarantee    2.263e+02  7.214e+01   3.137 0.001757 **
Sport_Model      2.811e+02  8.226e+01   3.417 0.000659 ***
Airco            2.430e+02  8.500e+01   2.859 0.004334 **
Tow_Bar         -2.203e+02  7.742e+01  -2.846 0.004523 **
Airbag_2        -2.167e+02  9.707e+01  -2.232 0.025847 *
Automatic        3.395e+02  1.428e+02   2.378 0.017586 *
Boardcomputer   -1.929e+02  1.126e+02  -1.713 0.087054 .
Power_Steering  -6.486e+02  2.715e+02  -2.389 0.017075 *
Airbag_1         4.773e+02  2.521e+02   1.893 0.058598 .
Doors            5.867e+01  3.958e+01   1.482 0.138578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1056 on 984 degrees of freedom
Multiple R-squared:  0.9121, Adjusted R-squared:  0.9103
F-statistic: 510.6 on 20 and 984 DF,  p-value: < 2.2e-16
```



R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 2: 후진소거법

- 모든 변수를 사용한 모델에서 중요하지 않은 변수를 하나씩 제거

```
73  # 변수선택 2: 후진소거법
74  backward_model <- step(full_model, scope = list(upper = as.formula(tmp_xy), lower = Price ~ 1), direction="backward", trace=1)
75  summary(backward_model)
76  anova(backward_model)
77
78  # 각 단계에서 제거된 변수 표시
79  backward_model$anova$Step
80
81  # 제거된 변수에 따른 AIC 감소분 표시
82  plot(backward_model$anova$AIC, pch = 15, cex=2, main = "AIC Decrease (Backward Selection)", xlab = "Number of Steps", ylab = "AIC")
83  text(backward_model$anova$AIC, backward_model$anova$Step, cex=1, pos=3, col="red")
```

```
> backward_model$anova$Step
 [1] ""                  "- CNG"             "- Cylinders"       "- Mfg_Year"        "- Mistlamps"
 [6] "- ABS"             "- Backseat_Divider" "- Central_Lock"    "- Met_Color"       "- Metallic_Rim"
[11] "- cc"              "- CD_Player"       "- Radio"           "- Radio_cassette"  "- Gears"
[16] "- Diesel"
```

R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 2: 후진소거법

- 변수선택 결과 (최초 36개 변수 → 21개 변수 선택 (15개 변수 제거))

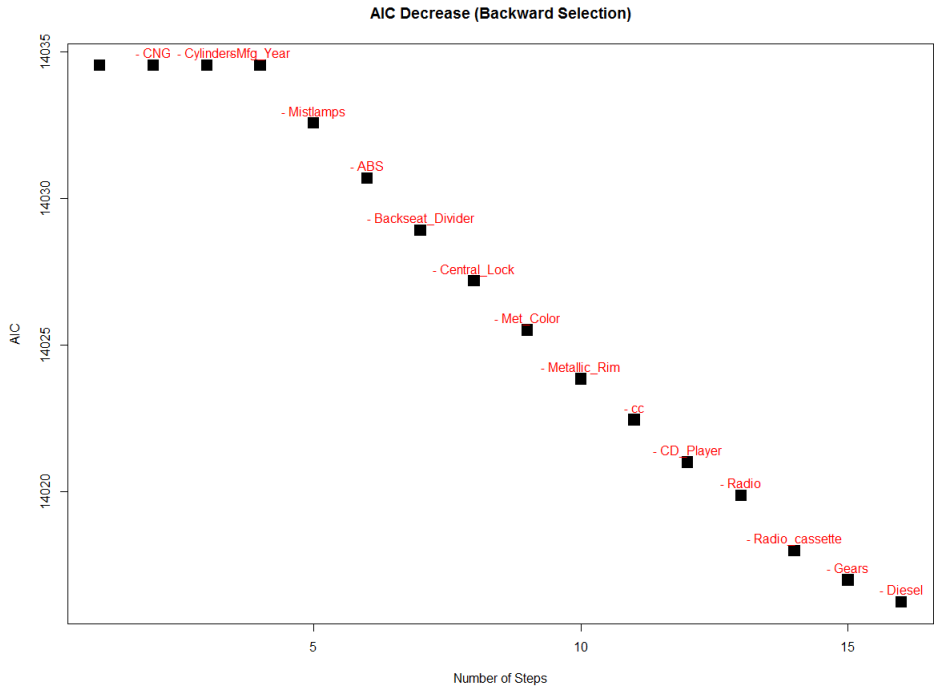
```
> summary(backward_model)

Call:
lm(formula = Price ~ Age_08_04 + Mfg_Month + KM + HP + Automatic +
    Doors + Quarterly_Tax + Weight + Mfr_Guarantee + BOVAG_Guarantee +
    Guarantee_Period + Airbag_1 + Airbag_2 + Airco + Automatic_airco +
    Boardcomputer + Powered_Windows + Power_Steering + Sport_Model +
    Tow_Bar + Petrol, data = trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6744.2  -643.7   -43.5   630.5  5924.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.528e+03  1.309e+03   3.460 0.000563 ***
Age_08_04       -1.176e+02  3.644e+00 -32.281  < 2e-16 ***
Mfg_Month       -1.067e+02  1.021e+01 -10.456  < 2e-16 ***
KM              -1.711e-02  1.327e-03 -12.898  < 2e-16 ***
HP               1.809e+01  3.236e+00   5.590 2.95e-08 ***
Automatic        3.373e+02  1.428e+02   2.363 0.018341 *
Doors            5.612e+01  3.965e+01   1.415 0.157290
Quarterly_Tax    1.156e+01  2.015e+00   5.738 1.27e-08 ***
Weight           9.259e+00  1.166e+00   7.938 5.57e-15 ***
Mfr_Guarantee    2.249e+02  7.215e+01   3.117 0.001879 **
BOVAG_Guarantee  4.138e+02  1.269e+02   3.260 0.001150 **
Guarantee_Period 7.511e+01  1.381e+01   5.437 6.82e-08 ***
Airbag_1         4.597e+02  2.526e+02   1.820 0.069067 .
Airbag_2        -2.272e+02  9.758e+01  -2.329 0.020071 *
Airco            2.377e+02  8.514e+01   2.791 0.005351 **
Automatic_airco  2.455e+03  1.746e+02  14.060  < 2e-16 ***
Boardcomputer   -2.056e+02  1.133e+02  -1.816 0.069700 .
Powered_Windows  4.620e+02  8.304e+01   5.563 3.41e-08 ***
Power_Steering  -6.192e+02  2.729e+02  -2.269 0.023479 *
Sport_Model      2.717e+02  8.273e+01   3.284 0.001059 **
Tow_Bar         -2.156e+02  7.754e+01  -2.780 0.005531 **
Petrol           7.055e+02  2.988e+02   2.361 0.018404 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1056 on 983 degrees of freedom
Multiple R-squared:  0.9122, Adjusted R-squared:  0.9103
F-statistic: 486.4 on 21 and 983 DF,  p-value: < 2.2e-16
```



R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 3: 단계적 선택법

- 상수항만 존재하는 모델에서 다음 절차를 번갈아가며 수행
  - ✓중요한 변수를 하나씩 추가
  - ✓중요하지 않은 변수를 하나씩 제거

```
85  # 변수선택 3: 단계적 선택법
86  stepwise_model <- step(lm(Price ~ 1, data = trn_data),
87                         scope = list(upper = as.formula(tmp_xy), lower = Price ~ 1), direction="both", trace=1)
88  summary(stepwise_model)
89  anova(stepwise_model)
90
91  # 각 단계에서 선택/제거된 변수 표시
92  stepwise_model$anova$Step
93  stepwise_model$anova$AIC
94
95  # 제거/선택된 변수에 따른 AIC 감소분 표시
96  plot(stepwise_model$anova$AIC, pch = 19, cex=2, main = "AIC Decrease (Stepwise Selection)", xlab = "Number of Steps", ylab = "AIC")
97  text(stepwise_model$anova$AIC, stepwise_model$anova$Step, cex=1, pos=3, col="black")
```

```
> stepwise_model$anova$Step
 [1] ""                  "+ Mfg_Year"         "+ Automatic_airco"  "+ Weight"           "+ KM"               "+ Powered_Windows"
 [7] "+ HP"              "+ Quarterly_Tax"    "+ Guarantee_Period" "+ BOVAG_Guarantee"  "+ Petrol"           "+ Mfr_Guarantee"
[13] "+ Sport_Model"     "+ Airco"            "+ Tow_Bar"          "+ Airbag_2"         "+ Automatic"        "+ Boardcomputer"
[19] "+ Power_Steering"  "+ Airbag_1"         "+ Doors"
```

R 실습: 다중선형회귀분석 및 변수선택

❖ 변수선택 3: 단계적 선택법

▪ 변수선택 결과

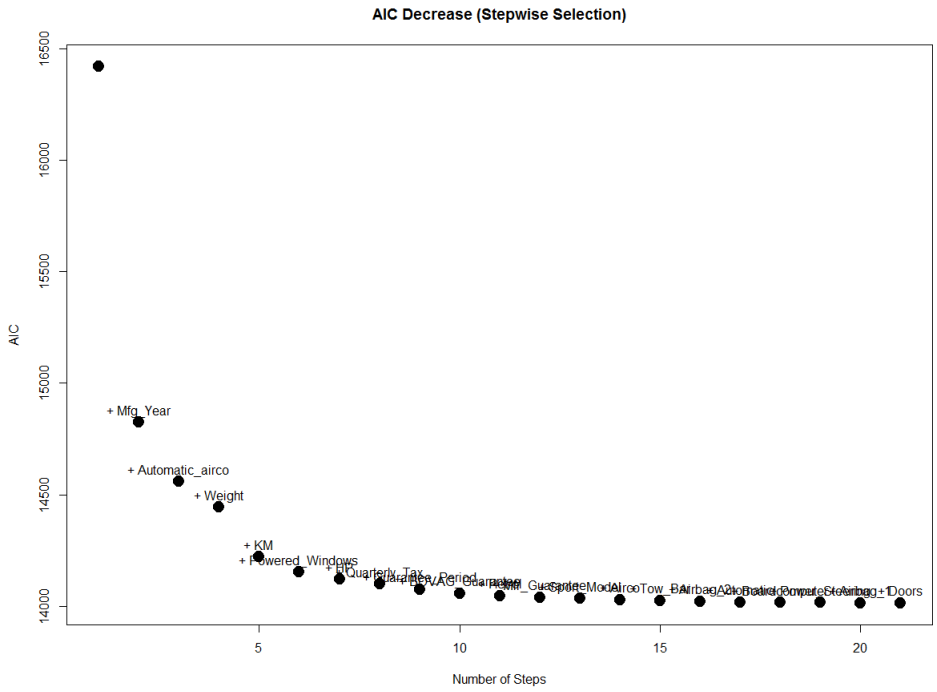
```
> summary(stepwise_model)

Call:
lm(formula = Price ~ Mfg_Year + Automatic_airco + Weight + KM +
    Powered_Windows + HP + Quarterly_Tax + Guarantee_Period +
    BOVAG_Guarantee + Petrol + Mfr_Guarantee + Sport_Model +
    Airco + Tow_Bar + Airbag_2 + Automatic + Boardcomputer +
    Power_Steering + Airbag_1 + Doors, data = trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-6747.2  -653.8   -53.8   640.8  5908.7

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.807e+06  8.542e+04 -32.857  < 2e-16 ***
Mfg_Year         1.402e+03  4.286e+01  32.718  < 2e-16 ***
Automatic_airco  2.451e+03  1.746e+02  14.037  < 2e-16 ***
Weight           9.233e+00  1.166e+00   7.918 6.50e-15 ***
KM              -1.734e-02  1.309e-03 -13.252  < 2e-16 ***
Powered_Windows  4.650e+02  8.300e+01   5.602 2.74e-08 ***
HP               1.819e+01  3.235e+00   5.625 2.42e-08 ***
Quarterly_Tax    1.146e+01  2.013e+00   5.694 1.63e-08 ***
Guarantee_Period 7.624e+01  1.377e+01   5.535 3.98e-08 ***
BOVAG_Guarantee  4.078e+02  1.268e+02   3.216 0.001342 **
Petrol           6.593e+02  2.956e+02   2.231 0.025933 *
Mfr_Guarantee    2.263e+02  7.214e+01   3.137 0.001757 **
Sport_Model      2.811e+02  8.226e+01   3.417 0.000659 ***
Airco            2.430e+02  8.500e+01   2.859 0.004334 **
Tow_Bar         -2.203e+02  7.742e+01  -2.846 0.004523 **
Airbag_2        -2.167e+02  9.707e+01  -2.232 0.025847 *
Automatic        3.395e+02  1.428e+02   2.378 0.017586 *
Boardcomputer   -1.929e+02  1.126e+02  -1.713 0.087054 .
Power_Steering  -6.486e+02  2.715e+02  -2.389 0.017075 *
Airbag_1         4.773e+02  2.521e+02   1.893 0.058598 .
Doors            5.867e+01  3.958e+01   1.482 0.138578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1056 on 984 degrees of freedom
Multiple R-squared:  0.9121, Adjusted R-squared:  0.9103
F-statistic: 510.6 on 20 and 984 DF,  p-value: < 2.2e-16
```



R 실습: 다중선형회귀분석 및 변수선택

❖ 각 변수선택 기법의 예측 정확도 비교 (검증 데이터 사용)

- MSE, RMSE, MAE, MAPE 네 가지 측면에서 비교

```
 99 # 검증 데이터에 대한 각 변수선택 결과의 예측 정확도 비교
100 full_haty <- predict(full_model, newdata = val_data)
101 forward_haty <- predict(forward_model, newdata = val_data)
102 backward_haty <- predict(backward_model, newdata = val_data)
103 stepwise_haty <- predict(stepwise_model, newdata = val_data)
104
105 # 회귀분석 예측성능 평가지표
106 # 1: Mean squared error (MSE)
107 perf_mat <- matrix(0,4,6)
108 perf_mat[1,1] <- mean((val_data$Price-full_haty)^2)
109 perf_mat[1,2] <- mean((val_data$Price-forward_haty)^2)
110 perf_mat[1,3] <- mean((val_data$Price-backward_haty)^2)
111 perf_mat[1,4] <- mean((val_data$Price-stepwise_haty)^2)
112
113 # 2: Root mean squared error (RMSE)
114 perf_mat[2,1] <- sqrt(mean((val_data$Price-full_haty)^2))
115 perf_mat[2,2] <- sqrt(mean((val_data$Price-forward_haty)^2))
116 perf_mat[2,3] <- sqrt(mean((val_data$Price-backward_haty)^2))
117 perf_mat[2,4] <- sqrt(mean((val_data$Price-stepwise_haty)^2))
118
119 # 3: Mean absolute error (MAE)
120 perf_mat[3,1] <- mean(abs(val_data$Price-full_haty))
121 perf_mat[3,2] <- mean(abs(val_data$Price-forward_haty))
122 perf_mat[3,3] <- mean(abs(val_data$Price-backward_haty))
123 perf_mat[3,4] <- mean(abs(val_data$Price-stepwise_haty))
124
125 # 4: Mean absolute percentage error (MAPE)
126 perf_mat[4,1] <- mean(abs((val_data$Price-full_haty)/val_data$Price))*100
127 perf_mat[4,2] <- mean(abs((val_data$Price-forward_haty)/val_data$Price))*100
128 perf_mat[4,3] <- mean(abs((val_data$Price-backward_haty)/val_data$Price))*100
129 perf_mat[4,4] <- mean(abs((val_data$Price-stepwise_haty)/val_data$Price))*100
130
131 # 변수선택 기법 결과 비교
132 rownames(perf_mat) <- c("MSE", "RMSE", "MAE", "MAPE")
133 colnames(perf_mat) <- c("All", "Forward", "Backward", "Stepwise", "GA_default", "GA_yourOwn")
134 perf_mat
```

```
> perf_mat
                All      Forward     Backward     Stepwise (
MSE  1.577365e+06 1.634343e+06 1.623485e+06 1.634343e+06
RMSE 1.255932e+03 1.278414e+03 1.274160e+03 1.278414e+03
MAE  9.121387e+02 9.242534e+02 9.211011e+02 9.242534e+02
MAPE 9.428209e+00 9.538071e+00 9.497384e+00 9.538071e+00
```

Q & A